Forthcoming in *Philosophy of Science*. ISSN: 0031-8248.

Title

Proof of Concept Research

(Draft Date: November 2019)

Author

Steve Elliott

Center for Gender Equity in Science and Technology

Arizona State University

P.O. Box 871108 Tempe, AZ, USA, 85287

stephen.elliott@asu.edu

Abstract

Researchers often pursue proof of concept research, but criteria for evaluating such research remain poorly specified. This paper proposes a general framework for proof of concept research that knits together and augments earlier discussions. The framework includes prototypes, proof of concept demonstrations, and *post facto* demonstrations. With a case from theoretical evolutionary genetics, the paper illustrates the general framework and articulates some of the reasoning strategies used within that field. This paper provides both specific tools with which to understand how researchers evaluate models in theoretical evolutionary genetics, and general tools that apply to proof of concept research more generally.

Acknowledgements

Thanks especially to Beckett Sterner, Norman Johnson, Kelle Dhein, Manfred Laubichler, and two anonymous referees for suggestions that substantially improved this paper. Thanks also to Norman Johnson and Adam Porter for insights into their project. Finally, thanks to Rick Creath and Patrick Forber for feedback on early versions of the ideas presented here.

1- Introduction

Proof of concept research is ubiquitous in science. Scientists commonly characterize early stage research as that of proof of concepts or proof of principles (e.g., Gould 2005; Servedio et al. 2014).[1] They employ it in many disciplines, from medicine and theoretical biology to engineering, and it can involve everything from exploratory work and testing causal hypotheses to data mining and computer simulation. Furthermore, funding agencies have developed programs for proof of concept research, though these often focus on translating scientific research into marketable widgets (e.g., NSF 2017). Given this heterogeneity of proof of concept research, how do scientists judge it as successful or not?

Philosophers have just begun to study and articulate general frameworks and evaluation criteria for proof of concept research. A promising route to do so is to study how scientists conduct and evaluate proof of concept research in practice (Plutynski 2005; Kendig 2016). This strategy aligns with similar efforts for studying and articulating evaluation criteria in kindred parts of science, such as in exploratory experimentation (Elliott 2007), data science (Elliott et al. 2016; Leonelli 2016), and simulation science (Winsberg 2015). By studying scientific practices and articulating these frameworks, philosophers better describe varieties of arguments and criteria that scientists use to evaluate their work and that of their peers. Once explicit, these criteria can be further studied, critiqued, and refined; and they can

[1] In this paper, I treat 'proof of concept' as synonymous with 'proof of principle', as is often done by researchers themselves (e.g., Schmidt 2006, 49).

be used to foster epistemological research into understudied institutions of science, such as funding agencies (O'Malley et al. 2009; Haufe 2013).

In this paper, I propose a general framework for proof of concept research, and I describe some of the arguments researchers use to evaluate proof of concept research as successful or not. Briefly, researchers who conduct proof of concept research aim to establish a prototype, and to argue for the continued development of that prototype in settings beyond those in which it was established. The "concept" proved in such research is that the prototype achieves functions or results expected or desired of it. The framework includes three primary parts: prototypes; proof of concept demonstrations, which establish that prototypes achieve their expected or desired results; and *post facto* arguments, which use heuristics to argue pragmatically for the further development of prototypes. This framework knits together and augments various aspects of previous discussions about proof of concept research so as to apply across the sciences. But the framework is not exhaustive. One route to develop it is to describe the varieties or kinds of prototypes, proof of concept demonstrations, and *post facto* arguments that scientists use in practice in many different fields. I do so here for theoretical evolutionary genetics.

Given longstanding debates about the value of theoretical evolutionary genetics and the results generated within it (Rao and Nanjudiah 2011; Okasha 2012), many have sought to account for how it yields good science. Recently, a team of theoreticians argued influentially that theoretical evolutionary genetics yields models that function as proofs of concepts (Servedio et al. 2014). That argument is consistent with Anya Plutynski's independently-developed account of population genetics as providing proofs of principles or of possibilities (2004; 2005; 2006a; 2006b). Here, I articulate some of the arguments used to evaluate prototype models in theoretical evolutionary genetics and show how they fit the general framework for proof of concept research. My discussion of these kinds of arguments isn't exhaustive, nor are the kinds of arguments articulated here always invoked, but they do reveal how biologists sometimes argue for theoretical evolutionary modelling as worthwhile both epistemically and pragmatically.

To develop and illustrate the general framework and articulate some evaluation criteria used in theoretical evolutionary genetics, I pursue the following strategy. In the next section I review contemporary discussions of proof of concept research, and in section 3 I knit them together and augment them in a general framework. Next I describe a case of a recent project that fits the kind of research characterized as proof of concept by Servedio et al. (2014). In section 5, I show how the case illustrates my accounts of prototypes and of proof of concept demonstrations. I also articulate two kinds of more specific demonstrations used in the case: coherence models calibration (CMC) and scope demonstration. In section 6, I articulate three heuristics used in *post facto* arguments, and I illustrate their use in the case study. Given the heterogeneity of proof of concept research, I don't claim that my discussion here exhausts it. Instead, I conclude by indicating further routes by which to develop the framework.

2- Contemporary Views of Proof of Concept Research

Among philosophers, Catherine Kendig has done the most to elucidate proof of concept research, illustrating it with cases from synthetic biology (Kendig 2016).[2] For Kendig, proofs are experiments with which researchers show that a hypothetical model, the concept, obtains in a test case. She discusses cases in which researchers first engineered Cyanobacteria to efficiently produce biofuels. These researchers engineered Cyanobacteria, they specified the causal hypothesis about producing biofuel as the focal concept, and they proved the concept with wet-lab experiments involving the bacteria. Kendig says that these experiments, taken as practices, shape knowledge-making categories, of which she highlights the concept of reengineering systems and processes, that enable generalization.

Kendig says that an experimental proof shows that a causal hypothesis is possibly true for a wider range of items like the object studied. After the experimental proof in the biofuel case, researchers classed the reengineered Cyanobacteria as good chassis organisms for further biofuel research. For Kendig, the general assumption distinctive of such experimental proof of concept research is that "*if it works here, it will also [work] in all cases like this*," when like cases share causal mechanisms, structures, functions, etc. (Kendig 2016, 740 [Kendig's italics]). She calls this a projectability assumption, and she indicates a

[2] Kendig distinguishes proof of concept research from proof of principle judgements, the latter of which are ethical judgments that supervene on the former, and will not factor into my discussion here.

somewhat broader version, arguing that successful test cases provide “justification in practice of the potential transportability of that research,” including both the concept proved and the practices used to do so (Kendig 2016, 737). This second assumption involves only potential transportability, whereas the first implies a stronger projection to all like cases.

Kendig’s discussion is a good starting point, providing several aspects for further development. First, the two projectability assumptions, and their application in cases, might be fruitfully elucidated to indicate how the assumptions are used differently. Second, it’s not clear how the projectability assumptions Kendig proposes differ from those used in other experimental projects not normally classed as proof of concept research. Consider situations in which synthetic biologists show how to use the already-established Cyanobacteria chassis to produce biofuel more efficiently, or with a slightly altered component set. Both of Kendig’s projectability assumptions apply in these situations, indicating that research that involves perhaps any tweak to a research design or object of study should be classed as proof of concept. If so, then a wider array of research than is often presumed should be classed as proofs of concepts. A different strategy would argue that there are some other aspects of proof of concepts research, beyond the application of the projectability assumptions, that more specifically characterize proof of concept research. I later pursue this second strategy by invoking the discipline-relative novelty of prototypes studied. Third, while synthetic biologists often employ computer simulations, Kendig focuses on wet-lab experiments. Her discussion can be complemented with accounts of how researchers evaluate and export computer models from other investigative contexts, especially more theoretical ones.

For theoretical investigations, Axel Gelfert argues that proofs of principles provide one criterion for evaluating the success of models (Gelfert 2016; 2018). He says that models function as proofs of principles when they establish that a "type of approach or methodology is able to generate potential representations of target phenomena," which he illustrates with Lotka-Volterra modelling, and he coins the useful term of "proof of principle demonstrations" to denote the things that do so (Gelfert 2016, 85). Ultimately, however, Gelfert's goal is to relate proof of principle research to his broader account of exploratory modelling, and he doesn't articulate forms for such demonstrations.

For evolutionary genetics in particular, Anya Plutynski argues that theoreticians provide proofs of principle or of possibility. When they demonstrate theorems, as did Fisher with his fundamental theorem of natural selection, they establish which conditions (e.g., population size, drift, etc.) influence a process (e.g., natural selection) that yields a general phenomenon (e.g., evolutionary change), and which conditions don't (e.g., orthogenesis, saltation) (Plutynski 2005; 2006b). When they analytically infer results from models, they prove "what must be so, for any population that fits (ceteris paribus) the description [of] the model," (Plutynski 2004, 1205). Either way, theoreticians explain which evolutionary phenomena are possible and/or necessary within a precise theoretical framework.

Plutynski notes that the role of theoretical models and of analytical results can change over time. Early in a research program, they can establish the explanatory scope of a framework (e.g., mathematized selection), they can obviate competing programs (e.g., orthogenesis), and they can show how to unify seemingly incompatible frameworks (e.g., Mendelism and biometry) (Plutynski 2004; 2005; 2006a; 2006b). Later, they might be used alongside other modelling strategies to provide a consilience of induction for a given evolutionary claim that otherwise lacks more direct evidential support, rendering that claim plausible (Plutynski 2001).

Plutynski's account suggests several routes for further development. Most importantly, she illustrates a philosophical program for articulating the wider variety of criteria that evolutionary geneticists use in practice to evaluate precision frameworks, theorems, and models. Two such criteria stand out for further articulation: how analytically-inferred results relate (1) to results inferred from theories couched in natural language, and (2) to results generated via computer simulation, now widely used among theoreticians, as cursorily discussed by Plutynski. Second, Plutynski showed that theoreticians also evaluate analytical results in relation to the prospects for competing research programs. Still to be articulated, however, are the structures of the arguments researchers use in making those assessments, especially in nascent projects.

Among biologists, a team led by Maria Servedio has recently proposed an influential framework for proof of concept research in theoretical evolutionary genetics (Servedio et al. 2014). They propose that a “clear verbal model lays out explicitly which biological factors and processes it is (and is not) considering and follows a chain of logic from these initial assumptions to conclusions about how these factors interact to produce biological patterns,” (Servedio et al. 2014, 2; compare with Plutynski 2006b, 61). But given that such processes are often complex, reasoning via verbal models alone is often too imprecise to settle disagreements about their soundness. Proof of concept models explicate the verbal model in a precise mathematical framework, make previously implicit assumptions explicit, and enable more precise and explicit chains of reasoning from models to consequences. By doing so, such models have a unique function. “The models *themselves* are tests of whether verbal models are sound; if their predictions do not match, the verbal model is flawed, and that form of the hypothesis is disproved (Servedio et al. 2–3 [Servedio et al.’s italics]).

This account has strengths and weaknesses. It provides a unique role for proof of concept research within theoretical evolutionary genetics, and it captures much of how theoreticians evaluate proof of concept models. But it also conflates concepts of models with that of chains of reasoning that invoke models, saying that mathematical models are *themselves* tests of verbal models. Furthermore, the account assumes that the only form of reasoning used to evaluate mathematical models as proofs of concepts is via comparison of predictions across verbal and analytical models. There is more to be said about the kinds of reasoning used in proof of concept research, especially for cases involving numerical computer simulations, now widely employed among theorists.

The accounts above miss many of the practical motives researchers have for conducting proof of concept research. Biomedical researchers are more explicit about these motives. For them, proof of concept research is as an aspect of Phases I or II in the standard IV-Phase evaluation process for therapeutic drugs and vaccines (Schmidt 2006). They note that such studies aim to roughly estimate treatment effect sizes so that researchers can design later-phase studies about the effectiveness of medicines (Gould 2005). Such studies also help researchers decide whether or not to proceed with further studies, and if successful, they enable researchers to recruit collaborators and secure funding for further studies (Schmidt 2006).

While all of those previous discussions are suggestive of a general framework for proof of concept research that applies across empirical and theoretical projects, none provides one. Given their focuses on particular sciences (experimental synthetic biology, theoretical evolution, medicine), they can appear narrow in scope. Furthermore, while they discuss the kinds of arguments and evaluation criteria used in proof of concept research, only a few have been sketched. In the next section, I knit together aspects from the previous discussions to address those two issues.

3- A General Framework

Proof of concept research aims to establish that prototypes achieve functions or results desired of them, and to argue for their continued development in settings beyond those in which they were established. A general framework for proof of concept research includes prototypes, proof of concept demonstrations, and *post facto* arguments.

Prototypes are artifacts that are importantly novel compared to their historical forerunners. This novelty, rather than the potential for exportability, distinguishes proof of concept research from later-stage research. Prototypes can be engineered physical items, and they can be theories or models. Forerunners are any items used prior to a prototype, but that may be similar in some sense to the prototype in structure or application, and that may provide pieces cannibalized and reused in the prototype. Researchers judge the novelty of prototypes relative to disciplinary norms. What those in one discipline consider novel, those in another might consider just a tweak to an extant artifact. Researchers evaluate prototypes with proof of concept demonstrations. Prototypes can be negatively evaluated, or fall short of the standards of a proof of concept demonstrations, and still be prototypes. They are just failed prototypes.[3] Kendig, Schmidt, and Gould focus on prototypes as physical artifacts like engineered bacteria cultures or medicines. Plutynski, Gelfert, and Servedio's team focus on models and theories. Regardless, they talk about novel artifacts at the leading edge of science.[4]

[3] The term *prototype* has a more neutral connotation than does Gelfert's (2016; 2018) phrase *proof of principle model*, and it avoids confusions invited by ambiguous modifications like *failed proof of principle models*.

[4] Prototypes can be other things not detailed here. Surgeries provide intriguing cases in which the prototype (the practice itself) and the method of demonstration might be usefully identified with each other. Prototypes might also be conceived in some cases as combinations of practices, physical artifacts, and theories; as when investigators develop new techniques to implant into a patient or animal model a novel device that will affect a hitherto speculative causal mechanism, say a brain implant to lessen narcolepsy. In such cases, neither the techniques, the device, nor the theory can be established as functional apart from each other. These complex prototypes provide interesting objects for further study.

Proof of concept demonstrations show that prototypes achieve functions or results desired of them. For physical artifacts, these demonstrations may be performances of a prototype staged for an audience, as with a launch of a novel rocket. More often these demonstrations are chains of reasoning or arguments that can be published. For physical prototypes, as in Kendig's biofuel case, these arguments relate data to descriptions of the prototype for empirical evaluation. For prototype theories or models, taken as abstract artifacts, these arguments show that the prototypes achieve epistemic functions or aims, such as describing, predicting, or explaining phenomena. Furthermore, those demonstrations may be modal, as with how-possible or how-actual explanations. In many cases, the kinds of arguments are not unique to early stage research. Scientists throughout the sciences and throughout different stages of research use arguments to show that theories and models perform epistemic functions. These arguments become proof of concept demonstrations only when they involve prototypes. In such cases, the phenomenon of interest may also be more circumscribed than in later stages of research. This account of proof of concept demonstrations augments Gelfert's (2016; 2018) discussion by explicitly invoking notions of chains of reasoning or arguments, by applying both in theoretical and in empirical investigations, and by acknowledging different modalities of phenomena as legitimate objects of study.

Finally, *post facto arguments* are arguments that researchers use after they have established a prototype with a proof of concept demonstration. *Post facto* arguments rely on fallible heuristics to argue for further research that uses the prototypes in question. It is these arguments, and their anticipated uses, that provide scientists with motivations for designing and conducting proof of concept research.

Post facto arguments license practical claims about how prototypes might further be used in new or expanded research projects. Researchers often design prototypes to be simple with the idea that, if they are positively evaluated, then they provide templates on which to build and evaluate more complex physical artifacts, models, or theories. Despite a demonstrated proof of concept, researchers often lack substantial justifications, characteristic of later-stages of research, for exporting these prototypes to further investigative contexts or for projecting early results to larger classes. This lack of substantial justification is due to the simplicity and novelty of prototypes, and to the fact that proof of concept research is often at the most distal leading edge of research, where relevant data and analyses are often scarce. *Post facto* arguments thus rely on fallible heuristics, like the projectability assumptions discussed by Kendig. They also underwrite arguments for comparatively evaluating competing research programs, as discussed by Plutynski. And they license practical issues of research, such as for expending further resources to develop prototypes or protocols, or conducting later stage research, as discussed by Gould and Schmidt.

This general framework for proof of concept research knits together and augments earlier insights about proof of concept research. The framework also enables a philosophical program of articulating the different kinds of proof of concept demonstrations and *post facto* arguments that scientists use in practice. In the sections to follow, I illustrate this program with a case study from theoretical evolutionary genetics.[5] Unlike previous discussions, I focus on a project that involves computer simulations. I show how the case illustrates each part of the framework, I articulate structures for proof of concept demonstrations heretofore undescribed, and I make explicit the kinds of heuristics the team used in *post facto* arguments. In so doing, the case study indicates how scientists sometimes judge results from theoretical population genetics as both epistemically and practically valuable.

4- Case Description

Constructing Multi-Model Systems

In 1998, Norman Johnson and Adam Porter at the University of Massachusetts, Amherst, began a project to study the genetics of speciation. They focused on speciation due to hybrid incompatibilities (HI), which they'd independently studied in animal populations. HI occurs when parents from related but distinct lineages yield hybrid offspring that are less fit than are the parents. Johnson and Porter felt that extant models for the genetics of HI failed to adequately capture how genes functioned in relation to each other, yielding unrealistic models and results. Johnson and Porter focused on epistasis, the phenomenon by which the products of multiple genes affect a single phenotype; and on pleiotropy, the

[5] The case exemplifies the kind of research discussed by Servedio et al. (2014), and the principle investigator has stated in personal communication that he conceptualizes the project along those lines.

phenomenon by which the products of a single gene affect multiple phenotypes. They thought that the mechanistic details of those phenomena could importantly affect HI, and that they could show how with quantifiable precision. I discuss the history of their project up to 2008.

Johnson and Porter (JP) began with a rough verbal model, in the sense of (Servedio et al. 2014), of the phenomenon they wanted to study. They later presented that model as a proposition for which they would study its plausibility. “We propose that regulated genetic pathways are a biologically realistic way to provide the complex epistatic gene interaction seen in empirical studies of hybrid fitness reduction. Here we investigate the plausibility of this proposition”, (Johnson and Porter 2000, 528). The quote indicates a verbal prediction: that gene regulatory processes, when varied across individuals within populations, can affect processes of HI and speciation. To quantifiably investigate the plausibility of that verbal model, JP developed a theoretical system of at least four models. Due to constraints on resources, JP opted to evaluate the multi-model system analytically and with individual-based computer simulations, rather than with actual populations. They ran their initial simulations in 1998 and 1999, publishing them in (Johnson and Porter 2000). The results netted them an NSF grant in 2000 to sustain their project into the mid 2000s.

The first model in the multi-model system, implicit in the verbal model, was a standard framework for how populations diverged from one another and speciated. In this framework, called the Bateson, Dobzhansky, Muller (BDM) model of speciation, two lineages of a founder population are prevented from interbreeding due to something like barriers or large distances between them. At regular intervals across generations, researchers enable a set of hybrids across the two lineages, and they compare the fitnesses of the hybrids to those of their parents. If the hybrids are less fit than their parents, then the researchers conclude that speciation is underway, and once hybrids aren't reproductively viable, the two lineages are distinct species. If JP could show that in BDM-like situations, selection on the mechanistic details of pleiotropy and epistasis could lead to HI, then they would show that those details influenced speciation processes.

The second model in the system was a mechanistic model for gene regulatory epistasis (Figure 1) that applied at the molecular level within organisms. Developmental geneticists usually study epistasis as a set of regulatory processes that include: genes producing RNAs and proteins; those RNA and protein products moving within and between cells; and those products attaching to target genes or putting in motion other molecules that attach to target genes. By attaching to target genes, molecules affect if, when, where in the organism, and how much those targeted genes produce their own products. Simple models often take the form of cartoons that represent a few genes, their products, the movements of those products, and their physical interactions.[6] JP introduced a concept of *binding strength*. This concept enabled a numerical assessment of fit between a product or transcription factor

[6] Models of interactions across dozens of genes are often in the form of wiring diagrams. These models involve enough genes that they are computably intractable within the framework of standard evolutionary genetics.

protein (tf) from one locus and a binding site or *cis* regulatory region on another locus. The stronger the bind, the more product the regulated gene produced. The concept of binding strength was roughly interpretable as a feature of molecular processes, and it was programmable for computer simulation.[7]

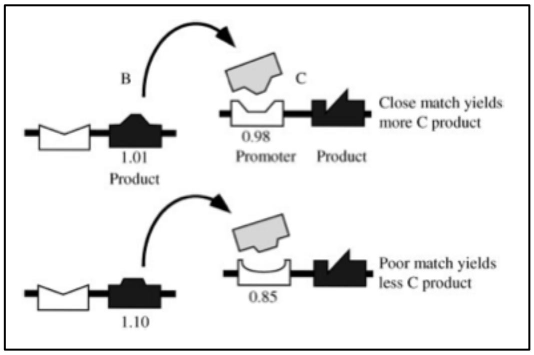
B
C
Close match yields more C product
1.01
Product
0.98
Promoter
Product
Poor match yields less C product
1.10
0.85

Figure 1: Gene regulatory pathway. The topmost interaction represents a close fit between the transcription factor from the allele at locus B and the *cis* region for allele C, yielding more products from C than in the bottommost interaction, which represents a looser fit. Reprinted with modifications from (Johnson and Porter 2000, Fig. 2).

The third model in the system was a set of three equations that JP wrote with the symbolic toolkit of evolutionary genetics. The first equation quantitatively defined binding

[7] The team developed a more biochemically realistic interpretation in their post 2008 work (Tulchinsky et al. 2014).

strength. The second equation defined an organism's phenotype as the product of all of the binding strengths in its regulatory pathway. Those two equations captured the mechanistic details of gene regulatory epistasis. The third equation defined an organism's fitness as the difference between its phenotypic value and the value of the optimal phenotype for its environment, which JP could define for pure analysis or simulation. With those three equations, JP had a precise model with which to study the potential for evolution within lineages due to variations in gene regulatory structures.

The fourth model was of organismal development and generational turnover. Organisms were hermaphrodites with few features, save for carrying the variable molecular interactions of interest. They were born, contributed gametes to a random gene pool, and died before the next generation was born. Each generation the pool of organisms faced two culling events, one random and one due to variation in fitnesses.

Evaluation Strategies

JP used at least three strategies to evaluate their multi-model system. First, they compared results from simulations using their multi-model system to results analytically inferred from a more-idealized version of their multi-model system. To create the more-idealized system, JP used the same four models described above, but they introduced several assumptions on the evolutionary genetic equations.[8] These assumptions enabled JP to derive probability distributions of expected fitnesses for hybrids of any given parental populations and any given generation. Next, they defined five discrete categories of fitness values (e.g., $W \geq 0.9$, or $0.1 \geq W \geq .01$), and they derived the expected probabilities for each category from the distributions. For example, the simplest case involved pathways of two genes each with one allele. After enough generations across infinite replications, half of the replicates would yield strongly unfit hybrids. For more complex regulatory structures, the proportion of strongly unfit hybrids would increase (Johnson and Porter 2000). These expected probabilities showed qualitative similarity with the predictions of the verbal model, and they became quantitative predictions for comparison with simulation results.

For their simulations, JP wrote a custom program that incorporated all of the models described above, performed selection on populations for thousands of generations, made hybrids, and calculated and recorded values for individual and population average fitnesses and phenotypes. For any set of starting conditions, JP ran dozens to hundreds of simulations,

[8] For instance, one assumption stipulated that at any given generation, each parental population exhibits exactly the mean optimal phenotype, regardless of the underlying genetic regulatory structures, which could differ across individuals and the two populations. In simulated or actual populations, parental populations at any given generation would, due to selection, approximate optimal phenotypes, but they wouldn’t exhibit *exact* optimal phenotypes.

also called replicates or runs. Then they collected and summarized the distributions of fitness and phenotype values across replicates, which they compared to the expected distributions derived from the more-idealized multi-model system.

For a second evaluation technique, JP compared the results of simulations with one set of starting conditions and parameter values against those of the same multi-model system but with different starting conditions or parameter values. JP systematically varied parameter values like mutation and migration rates; starting conditions like population size, complexity of regulatory pathways, number of generations; and the type of selection process, such as one-way directional, two-way directional, and stabilizing.

For a third technique, JP developed a multi-model system that shared the organismal and BDM models, but had a more standard (multiplicative) model of gene epistasis in place of JP's (regulatory) model. They compared the results of simulations on similar parameter values across those two multi-model systems.

Results

JP established several primary results. First, their simulated results were quantitatively similar, within 95% confidence limits, to their analytical predictions. For instance, for repeated runs on the simplest case, final hybrid mean fitnesses were very high (> 0.8) in 49.7 percent of runs and very low ($< 10^{-4}$) in 50.3 percent of runs, which was consistent with their analytical predictions of 50 percent for each (Johnson and Porter 2000). As the team increased the complexity of the regulatory model to include up to 10 loci and up to 2 alleles per loci, the proportion of hybrids with very low frequencies increased with the complexity of the regulatory model. Furthermore, simulations of the multi-model system that used the standard (multiplicative) model of epistasis yielded radically different results. In none of its runs did hybrids have fitnesses lower than 0.5, and in most they were higher than 0.8.

These results indicated that selection on phenotypes, which arose from regulatory molecular pathways, could lead to HI and speciation much more rapidly than could selection on phenotypes that developed due to genes that were linked via non-regulatory epistasis. JP modified their research system, and reached complementary results, for cases of occasional gene flow between parental populations (Porter and Johnson 2002), and for cases of pleiotropic gene regulation (Johnson and Porter 2007).

5- Prototypes and Proof of Concept Demonstrations

Prototypes are artifacts that are importantly novel compared to their historical forerunners. For JP, the prototypes were the verbal model and the precise multi-model systems that included the BDM, gene-regulatory, evolutionary-genetic equation, and organismal development models. Proof of concept demonstrations establish prototypes by showing that prototypes achieve functions desired of them. For JP, these demonstrations showed that the multi-model systems describe, predict, and explain abstract and general phenomena of rapid speciation. For the rest of this section, I articulate two kinds of proof of concept demonstrations used by JP: coherence models calibration and scope demonstration.

Coherence Models Calibration

Coherence models calibrations (CMCs) are arguments that show that a set of non-identical models or model systems yield results that are relevantly comparable, that the models are coherent with each other, and that they denote the same target systems. The general structure of CMCs is similar to that for what philosophers call derivational robustness, so it's important to clearly distinguish the two (Weisberg 2006; Woodward 2006; Wimsatt 2007; Kuorikoski et al. 2010). Both aim to infer something about two or more distinct models or multi-model systems by comparing the results derived from those disparate models, which we have no *a priori* reasons to believe will yield identical results. For derivational robustness, consider a model from which researchers analytically derive a result, and consider that the model is comprised partly by assumptions, some of which describe causal mechanisms or processes, but others of which introduce known falsehoods to enable the derivation. In such cases, researchers worry that the result may be due to the false assumptions. So they construct a second model, or more, with the same causal assumptions but different false assumptions that also enable the derivation. If researchers derive the same results or theorems from those additional models, then they conclude that theorems or results are robust to different assumptions.

Like derivational robustness arguments, CMCs compare different models that denote roughly the same phenomena or processes, they compare results derived from the different models, and they do so by considering those models under different assumptions. Unlike derivational robustness arguments, CMCs enable comparisons of models developed with different symbolic toolkits. They also enable comparisons of results characterized within those toolkits, even if those results were demonstrated with rules or techniques particular to those toolkits.[9]

JP's CMC, represented in Figure 2, is as follows. There are three multi-model systems, each of which is a prototype, designated as M_1, M_2, and M_3. And there are three outcomes, O_1, O_2, O_3, each of which is specific to its similarly indexed multi-model system. M_1 represents JP's verbal or natural language model, and O_1 is the outcome demonstrated via informal verbal reasoning from the model.[10] For JP, the verbal model encompasses their descriptions of gene regulatory pathways and speciation processes, and the relevant outcome is the claim that selection on the former can influence the latter, as shown by the quoted passage in the previous section. M_2 is the most-idealized multi-model system of the two

[9] Given these generalizations, it may prove useful to class derivational robustness demonstrations and CMCs as kinds of a more general form of argument, a project for further study.

[10] I take 'outcome' to denote a broad class of things producible from a model, such as a theorem or a set of simulated results. Similarly, I take 'demonstration' to denote a broad class of reasoning strategies that may include analytic derivations, abductive or inductive inferences, or simulated results. In this latter use, I follow Hughes's account of denotation, demonstration, and interpretation account of modelling (Hughes 1999).

precise systems JP created. Importantly, M_2 is an explication of M_1 in a mathematical framework capable of making precise assumptions and analytic inferences. O_2, which JP analytically derived from M_2, includes the expected probabilities for discrete fitness categories across infinitely possible replications. M_3 is JP's operationalization of M_2 and was programed into software to enable simulations of evolving populations. While it shares many idealizing assumptions of M_2, it lacks those that enable analytic inferences from M_2. To generate results from M_3, JP had to run their simulations, which yielded data about, among other things, fitness and phenotype values in parental populations and in hybrids. They then structured these data in data models and analyzed them across many replicates to yield proportions of replicates with speciation events, which comprise O_3.

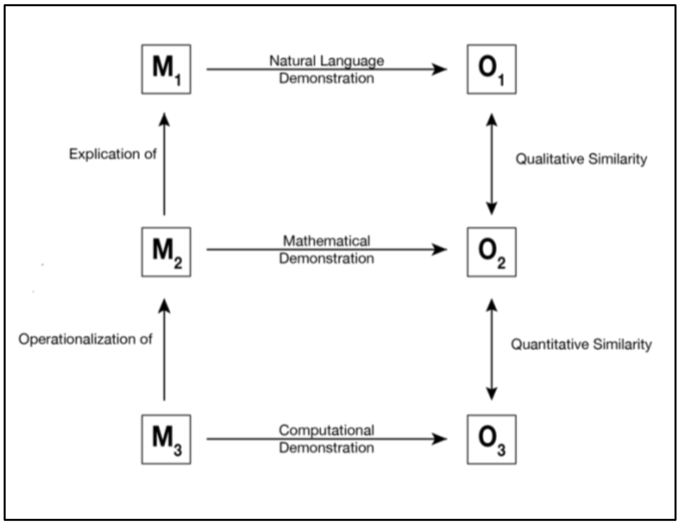

Figure 2: Johnson and Porter's Coherence Models Calibration.

Key to JP's evaluation of their CMC is the comparison of results, specifically O_2 to O_1 and O_3 to O_2. In Figure 2, I label the relevant relations as qualitative similarity and quantitative similarity, respectively. JP used an implicit and somewhat vague notion of qualitative similarity to compare O_2 to O_1. But they used explicit relations of quantitative similarity or predictive accuracy to compare O_3 to O_2.[11] Based on their verbal model (M_1), JP inferred that selection on gene regulatory networks could lead to hybrid incompatibilities (HI) and eventually speciation (O_1). Based on a highly idealized set of models couched in the mathematical language of theoretical evolutionary genetics (M_2), JP derived probability distributions for hybrid fitnesses. These showed that for the simplest systems, speciation would occur in half of the replicates, and more often in more complex systems (O_2). This outcome was qualitatively similar to that of the verbal model. After they operationalized those models for computer simulation (M_3), JP simulated biological systems and collected results (O_3) that were statistically significant and similar to those analytically inferred. Thus,

[11] These similarity relations are called prediction for several reasons. First, the analytic distributions are *conceptually* prior, in an inferential sense, to the tabulated distributions generated from the computer simulations, when both are used in a statistical test of similarity. Second, the analytic distributions can be calculated before, or *temporally* prior, to the tabulated distributions, thus establishing an expectation to be met or not in simulated systems. For the JP case, the first sense is more relevant than the second, but the ambiguity of 'prediction' is confusing enough for me to talk instead about quantitative similarity.

they showed that their M_3 could represent, predict, and explain possible speciation processes. Their CMC showed that M_1, M_2, M_3 are coherent with each other, that they denote the same target phenomenon or process despite being constructed with different symbolic toolkits.

Figure 2 and the discussion above capture much about reasoning practices in theoretical evolutionary genetics. The relations between M_1, M_2, O_1, and O_2 make explicit the relations and framework for purely formal work as discussed by Servedio et al. (2014) and to some extent Plutynski (2006b), and they clearly distinguish models from demonstrations. Furthermore, the relations between M_2, M_3, O_2, and O_3 indicate how simulation studies, now widely used in the field, fit into Servedio et al.'s approach. Thus, the account of CMCs builds on Servedio et al.'s account of proof of concept research while avoiding its limitations.

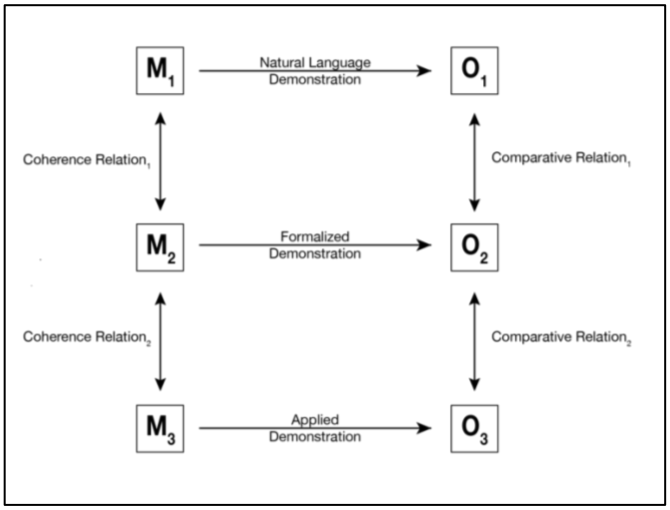

Figure 3: General Framework for Coherence Models Calibrations.

Figure 2 can be generalized. If used in other fields, the overall structure of models and outcomes might be the same, but due to different disciplinary norms and standards, the specific coherence relations between models and the specific comparison relations between outcomes might differ. These relations also depend on the kinds of questions pursued and systems studies. Figure 3 represents a more general structure for CMCs. For it, between models there are coherence relations, of which explicatory and operational relations are kinds. Furthermore, between outcomes there are comparative relations, of which qualitative and quantitative similarity relations are kinds. Demonstrations from M_3 to O_3 are more generally characterized as applied generalizations, and they can involve computer simulations, tests via empirical data collection, etc. Importantly, any given project may forego either analytic or applied demonstrations, or it may combine them. But to use a CMC, a project must compare at least results or outcomes demonstrated from at least two models or multi-model systems built with different symbolic toolkits.

In general a team establishes a prototype via a CMC by arguing as follows:

(1) Coherence relation$_1$ between M_1 and M_2 is met.

(2) Standards of inference are met for the Natural Language and Formalized/Analytic Demonstrations.

(3) Comparative relation$_1$ between O_1 and O_2 is met.

(4) Coherence relation$_2$ between M_2 and M_3 is met.

(5) Standards of inference are met for the Applied Demonstration.

(6) Comparative relation$_2$ between O_2 and O_3 is met.

Two inferences are implicit in proof of concept demonstrations via CMC.

(7) If (1), (2), and (3) are true; then M_1 and M_2 are mutually coherent, that is, they denote the same target system.

(8) If (4), (5), (6), and (7) are true; then M_1, M_2, and M_3 are mutually coherent, that is, they denote the same target system for at least the starting conditions, parameter values, and assumptions specified.

When researchers show that (1) through (8) are true, they show that a series of models built from different symbolic toolkits denote the same phenomena or target system. They are conceptually of a piece.[12]

Model calibrating denotes the activities of building an argument like (1) through (8), which can be represented in a graphic like Figure 3. In the course of a project, a team calibrates their package of models M_1 through M_3 to each other. They ensure that community or disciplinary standards are met. If any of those standards are violated, then a team must tinker with their package of models. They may adjust parameter values, model structures, or formalisms used to represent their models. In tough cases, they may question and adjust disciplinary standards.

[12] CMCs might be used to elucidate positions about the relative epistemic value of simulated results compared to empirical results (Winsberg 2015), a project for further study.

Scope Demonstration

Scope demonstrations are arguments that establish the range of phenomena to which a model or multi-model system applies. A *scope demonstration* builds on a previously employed demonstration, like a CMC, which is often relative to a singular set of starting conditions, parameter values, and other background assumptions. To determine the range of target systems or phenomena for which the models are coherent with each other, researchers systematically alter those conditions, values, and assumptions, and they re-compare the outcomes. This practice of systematic alteration has been discussed in literature related to exploratory experimentation (Elliott 2007), computer simulation (Winsberg 2015), and sensitivity analysis, a second kind of robustness analysis (Weisberg 2006; Woodward 2006; Wimsatt 2007). In a given project, a scope demonstration represents the results of those systematic alterations, and it argues for a range of phenomena to which the model applies. When the alterations are to real world systems, the conclusion is about the range of actual phenomena. When they are to theoretical systems, the conclusion is about the range of possible systems, as in JP's case.

JP employed two kinds of scope demonstration. First, they compared results generated across the same multi-model system but with altered causal complexity (e.g., 2 through 10 loci in gene regulatory pathways), different parameter values (e.g., mutation rate, migration rate), or simulation structures (e.g., number of replicates). An example of their results is represented in Figure 4. The figure compares the results generated from groups of replicates differing in the causal complexity of the model used. The top of the figure shows a frequency distribution of hybrid fitnesses when the regulatory pathway spans 2 loci, and the bottom shows a similar distribution when the regulatory pathway spans 4 loci. The comparison shows that the more complex the pathway, the higher the proportion of hybrids with low fitnesses. JP constructed dozens of these comparisons, some of which showed how much parameter values could be altered before there would be effects on hybrid fitness values.

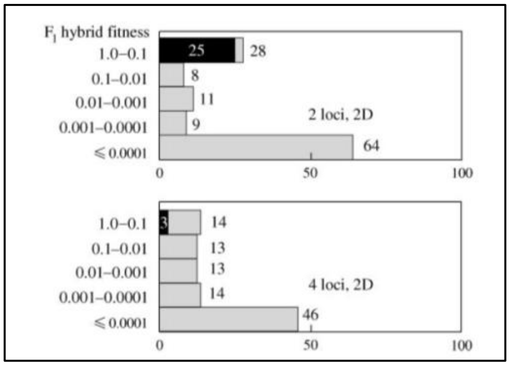

Figure 4. Frequency distributions of hybrid fitnesses for pathways with 2 and with 4 loci. Horizontal axis indicates number of replicates, vertical indicates fitness values discretized into 5 categories. Black bars indicate fitnesses above 0.5. All other parameter values and starting conditions held constant. The comparison indicates that the more complex the pathway, the higher the proportion of unfit hybrids. Reprinted with modifications from (Johnson and Porter 2000, Fig. 5).

For JP’s second kind of scope demonstration, they compared results generated across multi-model systems that differed not in degree of epistasis but in kind, and otherwise had identical parameter values, starting conditions, and simulation structures. In particular, they swapped out their gene *regulatory* mechanistic model and evolutionary genetic equations for a more traditional model and set of equations for *additive* epistasis. They reran their simulations and developed graphics, like those in Figure 4, for replicates using the traditional model. They then compared the new graphics to those developed for their regulatory multi-model system.

From both of these kinds of comparisons, JP established that, unlike the additive epistasis model system, their multi-model system could account for rapid speciations that happened in fewer than several thousand generations in a BDM scenario. These results didn’t confirm their multi-model system or disconfirm the more traditional account of additive epistasis. Rather, they showed the types of target systems to which each applied. As JP continued their project over the years, scope demonstrations became increasingly important to them, and they devised increasingly sophisticated means of comparing simulated results (Porter and Johnson 2002; Johnson and Porter 2007, Tulchinsky et al. 2014).

6- *Post Facto* Arguments

Researchers use *post facto* arguments to license further research beyond the context in which a prototype was established. *Post facto* arguments employ heuristics in Wimsatt's sense of non-truth preserving algorithms used by agents of bounded rationality, especially in contexts of transforming intractable problems into tractable ones. Heuristics typically rely on imperfect or false models of the world that sometimes fail in systematic ways (Wimsatt 2007, Chs. 5 & 6). In this section, I describe three heuristics that scientists use in arguments after they have established prototypes, illustrating them with JP's case.

The first heuristic involves the portability of the models to further research contexts and phenomena of study.

(H1) If a proof of concept has been demonstrated (e.g. (7) and (8) are true), then the theory, model, or package of models used in the demonstration (e.g. M_1, M_2, and M_3) can be used to study phenomena or target systems for which justifiably similar assumptions and parameter values apply.

(H1) accommodates Kendig's (2016) account of the transportability of prototypes and generalizes it to include proof of concept demonstrations in a wider array of investigative contexts beyond wet-lab experiments. Furthermore, when coupled with a scope demonstration, (H1) can be generalized to address a range of parameter values, assumptions, and starting conditions from a scope demonstration.

I locate arguments that use (H1) not as part of the proof of concept demonstration itself, as Kendig seems to, but as *post facto* arguments. Doing so enables us to distinguish the epistemic success of a proof of concept demonstration from the dispersive effects of such demonstrations. Not all prototypes are exported to further research contexts, even if they have successfully proved a concept. Drawing this distinction also saves us from having to treat all cases of science that employ a heuristic like (H1) as cases of proof of concept research. As a heuristic, (H1) can be tweaked to excise talk of 'proof of concept', and then it can be seen as a special case of a more general heuristic of empiricism.

(H1) invokes the phrase "justifiably similar" to underwrite the scope of exportability. This phrase is important for several reasons. First, those reasons that justify similarity are context-dependent. The kinds of reasons that researchers give for exporting models in one discipline may differ quite a bit from the kinds of reasons given in a different discipline. For any given discipline, those local norms must be described to better understand if and how these heuristics are applied, and how they may mislead. Second, and more importantly, the phrase enables a place for relations between simple and more complex models that target the same phenomenon, a feature central to the rationale of many proof of concept projects but often overlooked by philosophers. For researchers, many phenomena are too complex to be understood with extant models, so they adopt an iterative approach to achieve understanding (Elliott 2012). They begin by building the simplest model with which they can still account for some key features of the complex phenomenon they are studying. When they can show that those models in fact account for those key features, that is—they provide a proof of concept demonstration for a novel model, then they assert that the models can be exported not just for understanding similar features in related phenomena, but for studying further features of the original complex phenomenon. In such cases, the prototype model functions as a template, on which researchers can add and subtract elements to account for increasingly complex phenomena.

In short, the phrase "justifiably similar" enables models to be exported so as to apply to analogous phenomena in different target systems (e.g. using HI models in different genera or using predator/prey models for cycles of economic growth), and to more complex systems of which the accounted-for phenomena provide but one set of features. This latter export, from simple systems to more complex ones, is a common reason for doing proof of concept demonstrations. JP relied on (H1) interpreted in this way to further develop their multi-model systems to include not just epistatic gene regulatory motifs, but also migration between organismal populations (Porter and Johnson 2002), pleiotropic gene regulatory motifs (Johnson and Porter 2007), and biochemically realistic interactions between genes and molecules (Tulchinsky et al. 2014). Regardless, JP exported their models and results only to scenarios of BDM-like speciation, which was for them one condition for successful projection.

A second heuristic focuses not on the models themselves, but on the practices or tools used to create them.

(H2) If a proof of concept has been demonstrated, then the tools or practices used to

> construct and evaluate the relevant models or theories might fruitfully be used in justifiably similar research efforts.

Researchers often look to proof of concepts demonstrations not for the models themselves, but for the practices, strategies, or tools used to construct and evaluate those models. As with (H1), proof of concept demonstrations aren't necessary to heuristically argue to export practices or tools, but they are often sufficient. The JP case used many practices and strategies that, though not unique to the project, the team suggested others adopt (Johnson and Porter 2001). These include using individual-based simulations to operationalize and evaluate analytic theories, quantitatively evaluating speciation hypotheses, and making biologically realistic models of gene interactions and pleiotropy.

A third heuristic is about the allocation of resources for further use of the model. This heuristic is perhaps least appreciated by philosophers, but is in many ways central to the rationales for proof of concept projects. Those who discuss proof of concept projects in medical research note that establishing the effectiveness of a treatment, say a drug, is a process fraught with issues beyond the drug's efficacy for ameliorating symptoms (Gould 2005; Schmidt 2006). There are also issues of identifying and weighing the impacts of side-effects, of ethical responsibilities to test subjects, and of the resources needed to further establish the efficacy of the drug. Issues of resource allocation are legitimate constraints on all scientific projects, including those in theoretical evolutionary genetics. If researchers are to make progress on understanding a complex phenomenon, such as rapid allopatric speciation, then they must sort through the various strategies and tools for modelling that phenomenon and find those that, given the range of actual alternatives, are most promising. Proof of concept studies provide useful (and socially rewardable) means for doing so.

> (H3) If a proof of concept has been demonstrated, then it is worth expending more resources on the development of the prototype model and the use of the strategies, practices, and tools used to construct the model.

The desire to use this heuristic drives much—maybe most—of proof of concept research. JP relied on something like it when they applied and received, based on their initial results, an NSF grant to continue their work. Funding agencies often rely on (H3) when choosing which of several proposals to fund. JP relied on (H3) to justify their own time and resources to further develop the model after their initial results.

In such cases, (H3) isn't the only reason that factors into decisions. When choosing between two models to further fund or develop, decision makers might note that both functioned adequately in proof of concept demonstrations, but that one is cheaper (cognitively or financially) to reproduce and redeploy, or that both represented phenomena, but only one was also predictively accurate. In these cases, additional heuristics are used in decision making. Regardless, (H3) provides a baseline heuristic.

7- Conclusion

This paper provides a general framework for proof of concept research, and it articulates the structures for some of the arguments used to evaluate such research. These results unify previous discussions of proof of concept research, and they further elucidate how researchers evaluate results from theoretical evolutionary genetics, an often contentious field, especially when pairing analytically derived theorems with computer simulations. These results also contribute to the program of articulating how researchers judge early stage science as successful or not.

This conceptual framework indicates directions for further epistemological research. There are surely other kinds of proof of concept demonstrations and *post facto* heuristics beyond those discussed here, and those discussed here aren't invoked in every case of proof of concept research. If further kinds are identified, articulated, and critiqued, then a wider array of research could become more clearly understood and perhaps be made financially respectable (O'Malley et al. 2009; Haufe 2013; Elliott et al. 2016).

A particularly fruitful example of a further heuristic is about theory integration (Mitchell 2003; Laubichler et al. 2018). Often, when researchers provide a proof of concept demonstration, they show their peers that achieving a desired theoretical aim, such as theory integration, can be accomplished and is worth further effort. Such theoretical aims are longstanding features of research in theoretical evolutionary genetics (Plutynski 2004; 2005). JP explicitly address theory integration in a 2001 review article, in which they feature their multi-model system and modelling practices as examples by which to further develop a mechanistic theory of adaptation (Johnson and Porter 2001, 45). With their proof of concept demonstration, JP cut through conceptual discussions about the (im)possibility of syntheses between developmental genetics and evolutionary genetics. Implicit is a fourth heuristic, such that if a proof of concept has been demonstrated with models from historically distinct theoretical frameworks, then researchers might fruitfully integrate more aspects of those frameworks to study hitherto intractable phenomena. This heuristic provides a promising object for further study.

References

Elliott, Kevin C. 2007. "Varieties of Exploratory Experimentation in Nanotoxicology." *History and Philosophy of the Life Sciences* 29: 313–36.

———. 2012. "Epistemic and Methodological Iteration in Scientific Research." *Studies in History and Philosophy of Science Part A* 43: 376–82. doi.org/10.1016/j.shpsa.2011.12.034.

Elliott, Kevin C., Kendra S. Cheruvelil, Georgina M. Montgomery, and Patricia A. Soranno. 2016. "Conceptions of Good Science in Our Data-Rich World." *BioScience* 66: 880–89. doi.org/10.1093/biosci/biw115.

Gelfert, Axel. 2016. *How to Do Science with Models: A Philosophical Primer*. New York: Springer.

———. 2018. "Models in Search of Targets: Exploratory Modelling and the Case of Turing Patterns." In *Philosophy of Science: Between the Natural Sciences, the Social Sciences, and the Humanities*, edited by Alexander Christian, David Hommen, Nina Retzlaff, and Gerhard Schurz, 245–69. Cham: Springer. doi.org/10.1007/978-3-319-72577-2_14.

Gould, A. Lawrence. 2005. "Timing of Futility Analyses for 'Proof of Concept' Trials." *Statistics in Medicine* 24: 1815–35. doi.org/10.1002/sim.2087.

Haufe, Chris. 2013. "Why Do Funding Agencies Favor Hypothesis Testing?" *Studies in History and Philosophy of Science Part A* 44: 363–74. doi.org/10.1016/j.shpsa.2013.05.002.

Hughes, R. I. G. 1997. “Models and Representation.” *Philosophy of Science* 64: S325–36. jstor.org/stable/188414.

Johnson, Norman A., and Adam H. Porter. 2000. “Rapid Speciation via Parallel, Directional Selection on Regulatory Genetic Pathways.” *Journal of Theoretical Biology* 205: 527–42. doi.org/10.1006/jtbi.2000.2070.

———. 2001. “Toward a New Synthesis: Population Genetics and Evolutionary Developmental Biology.” *Genetica* 112–113: 45–58.

———. 2007. “Evolution of Branched Regulatory Genetic Pathways: Directional Selection on Pleiotropic Loci Accelerates Developmental System Drift.” *Genetica* 129: 57–70. doi.org/10.1007/s10709-006-0033-2.

Kendig, Catherine Elizabeth. 2016. “What Is Proof of Concept Research and How Does It Generate Epistemic and Ethical Categories for Future Scientific Practice?” *Science and Engineering Ethics* 22: 735–53. doi.org/10.1007/s11948-015-9654-0.

Kuorikoski, Jaakko, Aki Lehtinen, and Caterina Marchionni. 2010. “Economic Modelling as Robustness Analysis.” *The British Journal for the Philosophy of Science* 61: 541–67. doi.org/10.1093/bjps/axp049.

Laubichler, Manfred D., Sonja J. Prohaska, and Peter F. Stadler. 2018. “Toward a Mechanistic Explanation of Phenotypic Evolution: The Need for a Theory of Theory Integration.” *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 330: 5–14. doi.org/10.1002/jez.b.22785.

Leonelli, Sabina. 2016. *Data-Centric Biology: A Philosophical Study*. University of Chicago Press.

Mitchell, Sandra D. 2003. *Biological Complexity and Integrative Pluralism*. Cambridge: Cambridge University Press.

NSF. 2017. “Program Solicitation: Innovation Corps - National Innovation Network Teams Program (I-CorpsTM Teams).” National Science Foundation 18-515. https://www.nsf.gov/pubs/2018/nsf18515/nsf18515.htm. (Accessed January 27, 2019).

Okasha, Samir. 2016. “Population Genetics.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2016. https://plato.stanford.edu/archives/win2016/entries/population-genetics/. (Accessed January 27, 2019).

O’Malley, Maureen A., Kevin C. Elliott, Chris Haufe, and Richard M. Burian. 2009. “Philosophies of Funding.” *Cell* 138: 611–15. doi.org/10.1016/j.cell.2009.08.008.

Plutynski, Anya. 2001. “Modeling Evolution in Theory and Practice.” *Philosophy of Science* 68: S225–36.

———. 2004. “Explanation in Classical Population Genetics.” *Philosophy of Science* 71: 1201–14. doi.org/10.1086/426773.

———. 2005. “Explanatory Unification and the Early Synthesis.” *The British Journal for the Philosophy of Science* 56: 595–609. https://doi.org/10.1093/bjps/axi124.

———. 2006a. “Strategies of Model Building in Population Genetics.” *Philosophy of Science* 73: 755–64. doi.org/10.1086/518631.

———. 2006b. "What Was Fisher's Fundamental Theorem of Natural Selection and What Was It for?" *Studies in History and Philosophy of Biological and Biomedical Sciences* 37: 59–82. doi.org/10.1016/j.shpsc.2005.12.004.

Porter, Adam H., and Norman A. Johnson. 2002. "Speciation despite Gene Flow When Developmental Pathways Evolve." *Evolution* 56: 2103–11.

Rao, Veena, and Vidyanand Nanjudiah. 2011. "J.B.S. Haldane, Ernst Mayr and the Beanbag Genetics Dispute." *Journal of the History of Biology* 44: 233–81. doi.org/10.1007/s10739-010-9229-5.

Schmidt, Bernd. 2006. "Proof of Principle Studies." *Epilepsy Research* 68: 48–52. doi.org/10.1016/j.eplepsyres.2005.09.019.

Servedio, Maria R., Yaniv Brandvain, Sumit Dhole, Courtney L. Fitzpatrick, Emma E. Goldberg, Caitlin A. Stern, Jeremy Van Cleve, and D. Justin Yeh. 2014. "Not Just a Theory—The Utility of Mathematical Models in Evolutionary Biology." *PLOS Biology* 12 (12): e1002017. https://doi.org/10.1371/journal.pbio.1002017.

Tulchinsky, Alexander Y., Norman A. Johnson, Ward B. Watt, and Adam H. Porter. 2014. "Hybrid Incompatibility Arises in a Sequence-Based Bioenergetic Model of Transcription Factor Binding." *Genetics* 198: 1155–66. https://doi.org/10.1534/genetics.114.168112.

Weisberg, Michael. 2006. "Robustness Analysis." *Philosophy of Science* 73: 730–42. https://doi.org/10.1086/518628.

Wimsatt, William C. 2007. *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality*. Cambridge, MA: Harvard University Press.

Winsberg, Eric. 2015. “Computer Simulations in Science.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2019. https://plato.stanford.edu/archives/spr2019/entries/simulations-science. (Accessed June 24, 2019).

Woodward, Jim. 2006. “Some Varieties of Robustness.” *Journal of Economic Methodology* 13: 219–40.