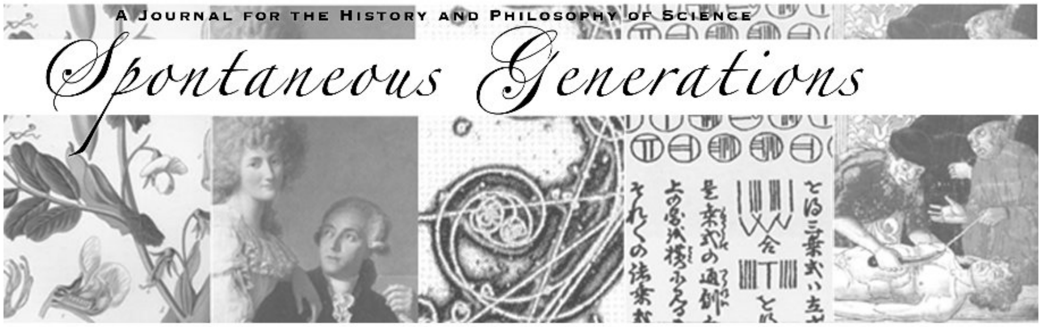

Going Outside the Model: Robustness Analysis and Experimental Science

Author(s): Michael Bycroft

Source: *Spontaneous Generations: A Journal for the History and Philosophy of Science*, Vol. 3, No. 1 (2009) 123-141.

Published by: The University of Toronto
DOI: 10.4245/sponge.v3i1.6118

EDITORIAL OFFICES

Institute for the History and Philosophy of Science and Technology
Room 316 Victoria College, 91 Charles Street West
Toronto, Ontario, Canada M5S 1K7
hapsat.society@utoronto.ca

Published online at jps.library.utoronto.ca/index.php/SpontaneousGenerations
ISSN 1913 0465

PEER-REVIEWED

Going Outside the Model
Robustness Analysis and Experimental Science*

Michael Bycroft†

In 1966 the population biologist Richard Levins gave a forceful and influential defence of a method called "robustness analysis" (RA). RA is a way of assessing the result of a model by showing that different but related models give the same result. As Levins put it, "our truth is the intersection of independent lies" (1966, 423). Steven Orzack and Elliott Sober (1993) responded with an equally forceful critique of this method, concluding that the idea of robustness "lacks proper definition and that its bearing on the question of whether a proposition is true is highly problematic" (533). Replies to Orzack and Sober, from Levins (1993) and Weisberg (2006b), have rejected the idea that RA, on its own, can confirm the results of models. I argue that these replies have not properly addressed Orzack and Sober's real criticisms, which focus not on the role of empirical data in RA but on the problem of ensuring that the models used in RA are independent. By drawing on accounts of RA in experimental science, I argue that there is in fact a fallible but viable form of RA that can both confirm the results of models and incorporate empirical data. However, for reasons other than those of Orzack and Sober, I conclude that RA may be of limited use in model-based science, as the example of fishery biology can show.

In the context of modelling, Robustness Analysis (RA) is a way of assessing the result of a model by showing that different but related models give the same result. But what does the robustness of a result tell us about the result? Steven Orzack and Elliott Sober (1993), writing in the context of population biology, criticise the view that RA can confirm the results of a model; that is, the view that the robustness of a result serves as evidence of its truth. Richard Levins (1993) and Michael Weisberg (2006b) read Orzack and Sober's discussion as an accusation that RA is an illegitimate form of confirmation because it does not draw on empirical data. In response to this critique, Levins and Weisberg develop a version of RA that does not confirm robust results and does not draw on empirical

*Received May 2009. Revised paper accepted October 2009.

†Michael Bycroft completed a Masters in History and Philosophy of Science at the Institute for the History and Philosophy of Science and Technology at the University of Toronto in 2008. His research centres on the philosophy of robustness and the role of robustness in early modern science.

data. I argue (in Part I) that this move towards *internal* RA is misleading, since there is a plausible general description of RA that is consistent with the method being *external*, that is, both empirical and confirmatory. Moreover, Orzack and Sober's criticisms are directed not at the absence of empirical content in RA but at the difficulty of showing that a given set of models are sufficiently *independent* to confirm a robust result. Replies to Orzack and Sober have neglected external RA, and Orzack and Sober's concerns about independence have gone unchallenged.

To address these issues I describe (in Part II) coincidence RA, an account of RA that has been widely discussed in experimental science but rarely applied to model-based science. Coincidence RA, I hope to show, is an empirical form of RA that offers a fallible but viable confirmatory tool to model-based science. Moreover, it can overcome the problems of independence that Orzack and Sober identify (Part III). However, it is by no means clear that external RA can make a smooth transition from experimental science to a model-based science, as the example of fisheries biology suggests (Part IV).

I. RA in model-building: internal and external versions.

Levins (1966) advocated RA in an influential paper on model-building in population biology (the study of how the members of populations interact with each other, with other populations, and with their environments). Levins begins by noting the variety and complexity of biological phenomena and the reliance of population biologists on simplified models of their target systems (such as predator-prey systems, fish populations, and sets of alleles). Levins suggests that models can be partitioned into their "essential" parts or "core assumptions," and their "artificial assumptions" or the "simplifications, distortions, and omissions that are introduced to facilitate the analysis" (1966, 423; 1993, 554). The core assumptions are "essential" in the sense that they are true or expected to be true of the target system, and also in the sense that, for a certain range of behaviours, the behaviour of the model depends more strongly on the core assumptions that on the artificial assumptions.

Levins writes that "[t]here is always room for doubt as to whether [the] result depends on the essentials of a model or on the details of the simplifying assumptions" (423). Thus the aim of RA is to find out which parts of a model are responsible for a given result: to locate the origin of the result. This is achieved in the following way:

> ...we attempt to treat the same problem with several alternative models each with different simplifications but with a common biological assumption. Then, if these models, despite their

> different assumptions, lead to similar results we have what we call a robust theorem which is relatively free of the details of the model. Hence our truth is the intersection of independent lies. (Levins 1966, 423)

This passage is controversial, but the basic idea is clear. To assess a result of a model, the modeller varies the artificial assumptions of a model, holding the core assumptions fixed. If the result is "robust" over these changes in the model, it must depend on the core assumptions. Levins seems to add, in the final sentence of the passage, that "truth" of the robust result follows from its robustness. Later, in a reply to Orzack and Sober, Levins clarifies this point, noting that RA does not itself confirm the result, and nor does it require gathering empirical data. Empirical data is used, in standard ways, to confirm the core assumptions of the model, and the robust result is true only if the core assumptions are true. Levins expects this two-step procedure to appease Orzack and Sober, who (in Levins' words) "are worried that the robustness strategy seems to propose a way to truth independent of observation" (1993, 54).

Weisberg (2006b) gives more detail than Levins about RA and how it fits in to the practices of modellers, but his account of RA is the same as Levins' account in two salient respects. Firstly, Weisberg takes seriously Orzack and Sober's scepticism about non-empirical confirmation: those authors are not only "rightly sceptical" about non-empirical confirmation but, on Weisberg's reading, this scepticism is the main basis for their scepticism about RA (2006b, 732-33). Secondly, Weisberg's response to this scepticism is to develop a version of RA that is neither empirical nor confirmatory. Weisberg differs from Levins in his account of how modellers use RA to show that certain assumptions of a model, and not others, are responsible for a result of the model. But he repeats Levins' distinction between the use of RA to identify the responsible assumptions, and the use of other methods to confirm those assumptions. Weisberg summarises these two moves in the following passage:

> Regarding robustness analysis as a non-empirical form of confirmation [as Orzack and Sober do] is also an oversimplification, one that distorts the origin of the confirmatory weight attached to robust theorems. We are now in a position to see why: Robustness analysis helps to identify robust theorems, but it does not confirm them.[1] (2006b, 742)

[1]This summary of Weisberg may not do full justice to his views as they are expressed in Weisberg (2006b). He writes, for example: "If a sufficiently heterogeneous set of models for a phenomenon all have the common structure, then it is very likely that the real-world

Although these two moves help both authors to develop interesting versions of internal RA, they lead away from some interesting terrain. In particular, they deny the possibility of external RA: a version of RA that is both confirms robust results and draws on empirical data. And they distract attention (including critical attention) from the real grounds for Orzack and Sober's scepticism about RA.

The rejection of external RA would be warranted if external RA were incoherent, in the sense that the basic idea behind the method is inconsistent with RA being both empirical and confirmatory. Orzack and Sober are silent on this question. But Weisberg offers a reconstruction of Orzack and Sober's reasoning that may be taken as an attempt to show that external RA is incoherent. The reconstruction is based on the following description of RA:

> Robustness analysis consists in analysing a set of models M_1,..., M_n, and showing that, for every i, $M_i \rightarrow R$. (Weisberg 2006b, 732)

Weisberg suggests that this procedure is "non-empirical" (732). This conclusion follows trivially if "analysing a set of models" is taken to mean "only considering models themselves." Now, it may be reasonable to define RA so as to exclude external RA by definition. Perhaps such a definition would simply reflect a consensus among modellers that RA is not worth doing if it involves gathering empirical data. But this would be to ignore a plausible description of RA according to which external RA is not ruled out by fiat. Consider the following schema:

> Robustness analysis is an attempt to locate the origin of R by showing, for some set of models M_1,..., M_n, that $M_i \rightarrow R$ for every i.

By "locating the origin of R" I mean identifying some property, entity, relation, concept or structure that is responsible for R, whether causally responsible, conceptually responsible, or some combination of the two. This schema is a plausible extension of RA as described by Weisberg and Levins, both of whom describe how modellers use RA to locate the origin of a result. However this schema does not exclude the possibility that RA can, so to speak, "reach into the world," since it leaves open the questions of whether the origin of R is located in the model or the

phenomenon has the corresponding causal structure" (739). Here Weisberg seems to describe a version of RA that has confirmatory power. However the suggestion is not developed further. Weisberg (2006a) is consistent with the reading I give in the body of this paper.

world, and whether it is responsible for RA in a causal or a conceptual sense or a combination of the two. Granted, the schema does not mention any data-gathering steps as part of RA. And one might argue that, once each M_i is sufficiently well-defined, showing that $M_i \rightarrow R$, for each i, can be completed as a purely conceptual (perhaps purely mathematical) exercise. Nevertheless, there is plenty of room for empirical data to enter into the schema: in showing that the models are sufficiently independent, for example, and in lending empirical support to the models. The schema is a plausible description of RA and it does not support a blanket rejection of external RA.

The challenge to advocates of external RA is to fill out the schema by describing how RA uses empirical data and how it confirms robust results. It may be that in fact there do not exist any workable ways of filling out the schema in this way. The best way to find out whether they exist or not, however, is to examine the candidate accounts of RA and check whether they can perform both tasks. As the accounts of Weisberg and Levins illustrate, simply showing that RA be *accompanied* by empirical data or by confirmation is not enough to show that these RA itself is using the data or is confirming the result. A real defence of external RA needs to show that confirmation and the use of empirical data are intrinsic to the procedure; that is, intrinsic to an attempt to "locate the origin of R by showing that R is robust." In Part II of this paper I consider a version of RA that is usually considered to be an example of external RA, and describe how empirical data enters into the procedure and how it confirms robust results.

As noted above, Orzack and Sober do not claim that external RA is incoherent.[2] True, they do not endorse any versions of RA that are empirical and confirmatory. But their grounds for doubt are not an in-principle rejection of external RA but a belief that there is a specific check on the confirmatory power of RA: the difficulty of defining the property of independence for models, and the difficulty of finding sets of reliable models that are independent. Insofar as they endorse any instances of RA being applied to models in population biology, they do so because the models used in those instances are more clearly independent than the models used in the instances they do not endorse, *not* because they are more clearly empirical.[3] Hence the aim of Part IV is to tackle Orzack and Sober head-on by defending RA against the problems of

[2]Indeed, Orzack and Sober have very little to say about non-empirical confirmation, aside from noting (1993, 538) that Levins' first version of RA seems to instantiate it.

[3]"Robust theorems," Orzack and Sober write, "would be desirable if we could get them. Perhaps, for example, a fair meiosis explanation (Leigh 1986) and Fisher's (1930) explanation for a 1:1 sex ratio are "independent" enough that they can be said to provide a robust truth [i.e. a truth that is confirmed by virtue of its robustness]" (540).

independence that they cite. This moves the discussion away from internal RA, but towards a more complete assessment of Orzack and Sober's criticisms of external RA.[4]

II. External RA in experimental science.

A promising place to look for an account of external RA is the philosophy of experimental science, where a number of authors have discussed the nature and merits of different forms of RA.[5] Judging by published work on RA, the possibility of such borrowing has not often been seriously considered. Wimsatt (1981) places Levins' account of RA in the context of a much broader discussion about the role of RA in distinguishing "that which is regarded as ontologically and epistemologically trustworthy and valuable from that which is unreliable, ungeneralizable, worthless, and fleeting" in scientific practice (128). But his account of how RA works for models is not the same as the account of RA usually advanced by authors in the context of experimental science, and he does not explicitly discuss RA in the context of experiment. Wimsatt is a shared reference for authors writing on RA in both models and experiment, but otherwise there seems to be little exchange of ideas between the two strands of literature.

In this Part I describe a form of RA, which I call *coincidence RA*, that philosophers have frequently applied to experimental science. I consider whether or not this is an external form of RA, and consider an immediate objection that might arise to the application of coincidence RA to models. I argue that coincidence RA offers a fallible but viable form of external RA to modellers, setting the scene for an assessment (in Part III) of how coincidence RA can address Orzack and Sober's worries about independence.

Various authors have gone into various levels of detail about coincidence RA and its variants.[6] It will be useful to give a brief sketch of the argument for coincidence RA, for which there are four steps. The first step is to show that multiple techniques give rise to the same result about a target object. The "techniques" may be instruments alone, or a combination of data from instruments and theories about how the instrument works.[7] The second step is to infer from the convergence in

[4]It is worth noting that Orzack and Sober seem to agree with Weisberg and Levins, more or less, on the value of internal RA. They describe a version of internal RA and note that it "allows one to distinguish between assumptions [of a model] that are strong determinants of a prediction [of the model] and those that are weak ones" (1993, 540).

[5]Philosophers of experimental science who endorse RA include Salmon (1978; 1984), Urbach (1981), Cartwright (1983), Hacking (1983), Franklin (1984; 1986; 1990), Culp (1999) and Weber (2005).

[6]Notably Salmon (1979; 1984), Culp (1995), Staley (2004), and Weber (2005).

[7]I use "instrument" in an expansive sense here to refer to the concrete parts

the results of the techniques that the results must have a common origin; that is, there must be some conceptual or concrete feature shared by the techniques that is responsible for this convergence. This step has been supported by an inference to the best explanation (for example, in Weber [2006, 283]) and by probabilistic arguments (for example, in Salmon [1984] and Culp [1995]). The third step is to observe that the common origin could lie *either* in a feature of the two techniques themselves, *or* in the target object that the techniques are designed to study. Now, if the techniques do not share any common features, the latter option is the only one available; that is, an object, property, or relation in the world is responsible for the common result. The property of "not sharing any common features," which the techniques are required to have, is usually called *independence*. The fourth step of coincidence RA is to infer that, since the target object is responsible for the common result, the result must be true of that target object.

One could imagine a fifth step in which it is inferred that the techniques used in the procedure must be reliable. After all, if the result is true, each technique has given a true result. This step is on sure ground, since showing that a technique reproduces a "test result"–a result that is known in advance to be true–is a widely accepted method of showing that the technique is more generally reliable.[8] Usually the truth of the test result is established by some other technique that is already known to be reliable, such as when a telescope is assessed by using it to view objects whose details we can observe with the naked eye. Whether or not coincidence RA is the origin of our confidence in the test result does not make any difference to attempts to assess techniques by the "test result" method; all that matters, as far as this method is concerned, is the degree of confidence in the test result.

Although this fifth step is warranted, it is not intrinsic to coincidence RA, precisely because it "does not make a difference" whether the truth of the test result is established by coincidence RA or by some other technique. Any other test result will do. So coincidence RA only has a special role to play in assessing techniques if there are few or no test results available by other means. If test results are available without the use of coincidence RA, then those test results (and not RA) can be used to assess the techniques that are in doubt. Of course, a single test result can be used to assess

of the technique that causally link the target object to observed data–even if those causal processes are not packaged in the form of stable, self-contained devices like microscopes. A container of smoke particles in Brownian motion is part of the "instrument" used to find Avogadro's number, for example.

[8]For example, Franklin (1986; 1990) gives various examples of this method being applied to experimental physics.

multiple techniques. This scenario mimics RA, but it is just a case of an ordinary method being applied multiple times.

Anyone who is sceptical about external RA will want to know how coincidence RA makes use of empirical data. It does so in a number of ways. The first is to help develop the techniques such that they are plausible; that is, they have a decent chance of giving consistently true results about a phenomenon. The plausibility of the techniques has a pragmatic and a logical role. Pragmatically, there are a large number of instruments and assumptions that would, to a pre-scientific mind, seem equally promising as techniques for a given target object. Only a small number of these will be actually reliable, and it would be inefficient to perform RA on these techniques in a trial-and-error manner until a robust result appeared. Logically, we would be justifiably suspicious if this trial-and-error procedure regularly gave robust results. The unlikelihood of a robust result arising by chance from two independent techniques needs to be weighed against the unlikelihood of two reliable techniques, each relevant to the same target object, being developed by chance.

Another role for empirical data in coincidence RA is in making judgments about whether the techniques used in the method are independent. There are two extreme views about independence judgments that any convincing defence of coincidence RA needs to debunk: one view is overly optimistic and the other is overly pessimistic. The optimistic view raises doubts about non-empirical confirmation. It is the view that independence judgments about techniques used in RA can be made just by analysing the logical relations between the theories about the different techniques, without considering whether or not those theories are true.[9] Independence judgments are not the only point of entry for empirical data into RA (see the previous paragraph), but they are arguably the *main* point of entry. RA that denies this point of entry looks like it needs very little empirical input to do its confirmatory work; in other words, it looks too good to be true.

One reason the optimistic view might have intuitive appeal is that, if two independent theories about two techniques give the same result, the common origin of this result must lie outside the theories; it must lie in the real world. But the real world includes not just the target object but the instruments used to generate data about the object. So the common origin may lie in the instruments and not the target object. Unless one *knows* that

[9] Culp (1995, 450-54) seems to take this view, using a logical notion of independence (two theories are independent if "the truth and falsity of one is isolated from the truth and falsity of the other"). This suggests that, according to Culp, techniques can be shown to be independent just by analysing the theories used in those techniques, ie. without confirming those theories.

the instruments do not have any concrete features in common, one cannot rule out the possibility that the common origin of the robust result lies in the instruments. Hence the optimistic view about independence judgments is overly optimistic, and does not raise worries about coincidence RA being a non-empirical form of confirmation.

Coincidence RA clearly makes use of empirical data to establish the plausibility and independence of the techniques used in the method. But is it really the robustness of the result that confirms robust results, according to coincidence RA? The overly pessimistic view of independence judgments, if true, would cast doubt on the confirmatory weight of robustness. The pessimistic view is that a pre-requisite for making correct independence judgments about a set of techniques is to have a well-confirmed theory about the instrument used in each technique.[10] This raises the suspicion that most or all of the confirmatory power of coincidence RA comes from this prior knowledge of the reliability of the techniques used.

However, it is not necessary to be confident that two techniques are reliable in order to be confident that they are independent. Consider the flagship case of coincidence RA, Jean Perrin's multiple derivations of Avogadro's number.[11] Avogadro's number is the number of molecules in one mole of a substance (the number of molecules in 2 grams of hydrogen gas, for example). One of the techniques used to derive the number involves observing the Brownian motion of smoke particles in a known volume of gas, a technique that makes use of a formula for the kinetic energy of massive particles ($E = \frac{1}{2}mv^2$) to deduce the mass of the particles in the gas. Suppose, for arguments sake, that the energy formula was not well-confirmed at the time of the derivation: perhaps some physicists had reasons to believe that the coefficient and the power in the formula were not $\frac{1}{2}$ and 2 respectively, but $\frac{1}{3}$ and 3. RA would still be effective in this case, since it is not necessary to know the precise form of the energy equation to know that the motion of the smoke particles, and our theory about particulate motion, is independent of (for example) the electroplating of silver, the basis for another of Perrin's derivations. It is enough to know that the energy formula deals with the motions of particles, and that the amount of silver that is deposited in the plating process is not highly sensitive to the precise way that particles move. Hence the pessimistic view about independence judgments is mistaken, and it does not threaten

[10] Weber (2005, 286) takes this stance, saying that the "antecedently established reliability of both techniques" in RA is required before one can be confident that a robust result from those techniques is not an artefact.

[11] This example has been discussed by many authors. I draw on the accounts in Salmon (1978; 1984) and Cartwright (1983). The original account is Perrin ([1913] 1923).

the standing of coincidence RA as a procedure that can itself increase our confidence in the truth of a robust result.

So far my account of coincidence RA has been guilty of idealisation and abstraction, masking ongoing debates about the effectiveness of the method. The second step of RA is to infer from a common result that a common origin, responsible for the common result, must be present. This description abstracts many details, and there is debate about how this step should be filled in: for example, Salmon's attempt to use a probabilistic "principle of the common cause" is one variant under attack.[12] Idealisation enters my account in the assumption that sufficiently independent and plausible techniques, giving sufficiently convergent results, will be available for any given target object. There are good reasons to think that these high-quality resources will be in short supply in scientific practice.[13] Relatedly, it is possible to find numerous examples in experimental science of failed applications of RA: cases where RA has led to a conclusion that was either found later to be false, or was immediately trumped by a method that could make better use of the resources available at the time.[14]

In RA's defence, it is worth noting that rival and more traditional forms of justification in science (induction, abduction, hypothesis-testing, and so on) are also limited by the quality of the resources available in the context of their use. Also, few (if any) writers on RA have denied that RA can be a powerful form of empirical confirmation when high-quality resources are available. None of the authors cited here have denied, for example, that external RA plays a legitimate and powerful confirmatory role in the case of Jean Perrin's derivation of Avogadro's number. The consensus among philosophers of experimental science seems to be that external RA is a fallible but viable method in experimental science, one that can be used well sometimes but must be used with caution at all times. Insofar as Orzack and Sober (and Levins and Weisberg) deny that external RA is a viable method for modellers, the viability of the method in experimental science is a fact in need of explanation. In Part III I consider the problems with independence judgments that Orzack and Sober cite in their attack on external RA, and argue that those problems do not supply the needed explanation.

Before moving on, it is worth noting here one difference between experiment and modelling that might seem to, but in fact does not, stand in the way of a successful borrowing of external RA by modellers. In the experimental case it is possible for the target object to be responsible,

[12] See Weber (2005, 281-83).

[13] See Stegenga (forthcoming, 5-6) for a summary of this kind of concern about RA.

[14] Wimsatt (1981), Rasmussen (1993), and Hudson (1999) give examples of these two kinds of failure of RA.

in a very direct way, for a common result. Consider two different kinds of microscope, an optical microscope and an electron-transmission microscope, focused on a bacterial sample. If the microscopes are working correctly then the common result will be concrete: one can literally see the similar images in the eye-piece of one and the screen of the other. There may be differences between the two images (one will be greyscale and the other not, for example). There may also be differences between the common features of the images and the actual bacteria (in respect of scale, for instance): differences such that the significance of the images is not obvious without drawing on further assumptions about the working of the instruments. But it remains true that the two microscopes yield very similar results solely through the causal interactions between the sample and the instruments. This is quite different from the modelling case, where common results arise out of a conceptual system (such as a set of mathematical equations) that has no direct causal link to the system it is meant to describe.

This difference is relevant because some authors have suggested that the first step of coincidence RA (the inference from a common result to a common origin) is only permissible when the common origin is a *causally* responsible for the common result. Cartwright (1983) maintains that coincidence RA works only in cases where we "infer a concrete cause from a concrete effect," the reason being that in such cases "we assume that causes make effects occur in just the way they do, via specific, concrete causal processes" (84). Similarly, Salmon (1984, 223) maintains that, for his variant of coincidence RA to work, "there must be causal processes leading from the causal background [the target system] to the correlated effects [the common result]."[15] If these restrictions on coincidence RA hold, one might wonder whether the method is applicable in situations, like model-building, where common results do not arise except through a complex mixture of empirical and conceptual investigation.

It may be right to value causal RA over mixed RA. But mixed RA must have *some* value–indeed, considerable value–because applications of mixed RA are among the cases of RA that are widely regarded as successful. Jean Perrin's derivations of Avogadro's number is one such case. None of Perrin's 13 derivations yield Avogadro's number as a direct causal product of an experiment. In each case, experimental data is

[15]Both Salmon and Cartwright may be read as placing much weaker restrictions on the scope of coincidence RA than the restriction described above. Indeed, both are well aware that (to quote Salmon) "none of the methods used to ascertain the value of Avogadro's number does so exclusively on the basis of directly observable quantities without any appeal to auxiliary hypotheses" (1984, 224, fn4). Nevertheless, both authors could be read as casting doubt on mixed RA.

coupled with theory about the causal processes behind the observations to give a calculation of the number. Hence the mixed-ness of RA in the models case should not on its own prevent modellers from making as much use of RA as Perrin did.

III. INDEPENDENCE: A REPLY TO ORZACK AND SOBER.

Orzack and Sober maintain that the idea of robustness "lacks proper definition and that its bearing on the question of whether a proposition is true is highly problematic" (533). This scepticism about the confirmatory power of RA derives largely from their concern about the notion of the independence of models. They build their argument against external RA by distinguishing three variants of the method. These variants differ in terms of what is known about the models used. Modellers might know that at least one of the models is true; know that all of the models are false; or not know whether any of the models are true. The variant of RA I consider here is closest to, but not identical to, the third of Orzack and Sober's options. As suggested in Part II, coincidence RA is defensible in the experimental case when applied to a set of techniques, each of which is known to be plausible. Replacing "techniques" with "models" in the previous sentence gives the variant of RA that I shall now consider in light of Orzack and Sober's concerns about the notion of independence in RA.

The reason for making judgments about the independence of two or more models is clear: to ensure that a result that is robust across those models is not due to a false assumption held in common by the models. Orzack and Sober think it is much less clear (firstly) what the property of "independence" amounts to, and (secondly) how a set of models can be independent and still give a robust result. I argue that the first of these problems is real, but that it is negligible for our purposes because it is no more a problem for RA applied to models than for RA applied to experiments. The substance of the second problem is not entirely clear, and the best answer to the problem is to present a procedure by which modellers might develop independent models on which to perform external RA. I will sketch an example from climate science to illustrate this procedure.

Orzack and Sober consider two ways of defining the "independence" of two models: statistical independence and logical independence. Their concern about statistical independence is that "assigning probabilities to models is not a coherent notion" (1993, 539). This may be a genuine problem for external RA (and RA in general), but it does not explain why coincidence RA should be any less viable in model-based science than in experimental science: philosophers of experiment struggle to make sense of the idea of assigning probabilities to experimental techniques.

The existence of Bayesian accounts of RA in experimental science might suggest otherwise. But even a well-developed Bayesian account, such as that given by Allan Franklin, concedes that in practice it is very hard to measure these probablities (1984, 51-2; 1986, 125). Indeed, the Bayesian calculations in Franklin's account are only used to establish the soundness, in principle, of RA. When it comes to discussing particular cases of independence, Franklin resorts to qualitative accounts. For example, to establish that the bubble chamber and the spark chamber count as "independent" ways of observing subatomic particles, Franklin gives a loose statement of the theory about the two instruments: one works by "bubble formation in superheated liquids" and the other by "electrical discharge in ionized gases" (Franklin 1986, 125). The world of models lacks a rigorous and practicable account of independence, but so does the world of experiment.

Orzack and Sober's second worry about independence is that, as a matter of fact, the models that can be used in any given instance of RA are unlikely to be independent: "Generally speaking, competing models are not independent" (1993, 539). Orzack and Sober do not say what "competing models" are, or why they are typically dependent on each other. A natural reading, which I will consider here, is that "competing models" are "models of the same target system." Examples of models that compete in this sense are two models of predator-prey systems that differ only in whether the account for the density dependence of the population and the other does not; or models of evolution in a variable environment that differ only in whether they take the environment to vary discretely or continuously.[16]

Why might two models of the same target system be more likely to be dependent than two models that are each of a different target system? Intuitively, the facts about a target system constrain the number and range of models that can accurately describe the system. So perhaps those facts constrain the models so much that no two models of the same system can be independent. A natural way to challenge this view is to emphasize the freedoms that are available to model-builders when considering a target system. Internal RA relies on one such freedom: some parts of a target system can be strongly determinant of (if not all the behavior of a target system) at least some range of the behavior of a target system. Hence there are many models that are reliable with respect to that range of behavior: all the models that represent the structure that dominates the target system, but vary in respect of the non-dominant parts of the system. This kind of freedom does not make room for external RA, of course, because it still requires that each model of the system–if it is to reliably

[16]These examples are taken from Weisberg (2006b) and Levins (1966) respectively.

describe the same behavior–must have in common a representation of the dominant structure.

Levins suggests another degree of freedom that models (including models of the same system) enjoy: "The difference between legitimate and illegitimate simplifications depends not only on the reality to be described but also on the state of the science," where the "state of the science" means the "problems" that the science addresses (1966, 422). The difficulty with this defense is that "dealing with different problems" may just mean "taking an interest in a different class of results." For example, Levins describes one "problem" in early population biology: "Could weak natural selection account for evolutionary change?" (422). Plausibly, this "problem" just picks out a particular class of results (evolutionary change) of the target system (populations under weak natural selection). If this is the case then this degree of freedom allows core assumptions to vary across models of the same target system, but at the cost of robustness: the class of trustworthy results varies as the core assumptions do.

Perhaps there are no other degrees of freedom that can support the independence of competing models. Regardless, Orzack and Sober's concern about the non-independence of competing models is unsatisfactory. For they do not say why RA must use only competing models and not non-competing models. One might wonder whether non-competing models will have *too much* freedom: surely they need to be somewhat constrained in order to give convergent results; that is, results about the same object, property, or structure. Coincidence RA suggests an answer, inspired by experiment: the two models are of two different target systems that are each causally linked to the same object, property, or structure. It is worth illustrating this answer with a (very) schematic example of how coincidence RA might be applied to models of climate change.

The first task is to pick out a target object or property, a phenomenon to record or detect. This phenomenon needs to be relevant to climate change. It also needs to play a causal role in two independent causal chains that lead from the target property to phenomena we can easily observe. The target property might be the level of CO_2 concentration in the atmosphere, and the observable phenomena might be the frequency of acid rain events and level of global temperatures. For each of these observables there is a causal chain leading to it from the target property (CO_2 concentrations).

The second step is to use background knowledge and data about global climate to create a model, for each chain, that represents the causal chain leading from the shared quantity to the observable quantity. These models do not need to be well-confirmed, but they need to be plausible, as discussed in Part II.

Thirdly, use the models and the observed quantities to infer, for each causal chain, the concentrations of CO_2. The more we rely on our models to make these inferences, the more effective RA will be.

Fourthly, check to see whether each model gives the same result about concentrations of CO_2. If they do, then we can be confident that the result is true. We can also be confident that the models are reliable, insofar as they were deployed in inferring the target property.

This procedure may not have precisely the same form as the procedure followed by experimenters who want to measure an unobservable quantity or test their theory about their techniques. For one thing, it is an example of "mixed RA," as defined above: it resembles Perrin's derivations of Avogadro's number, not observations of the same target object made through different microscopes. But the *outcome* of the procedure is the same as in the experiment case: two techniques, based on two causal processes, that support an inference from the divergent products of those causal processes to their shared causal origin.

IV. Fishery biology: external RA in practice.

Orzack and Sober's doubts about the independence of models do not stand up to scrutiny. But nor does external RA stand up to the scrutiny of actual practice in model-based science. With a clear idea of how external RA works in ideal conditions, we are in a position to see why the method fails when conditions turn against it.

Consider the case of fishery biology. This is a good test case, since fishery biologists tend to develop relatively realistic and precise models of specific fish populations (Levins 1966, 422), the kind of models that might support external RA. A glance at the methods of this field shows that external RA is very rarely used, if at all. One might argue that external RA is used on a small scale. To take one example, electrofishing and gillnet fishing are techniques that can be used in tandem to give robust sets of data about lake fish populations (as discussed in Eros et al., 2009). And one might say that this small-scale RA involves models. But the models are simple pictures of (for example) how fish interact with nets, not models of the fish population and its large-scale dynamics. Models of the latter kind contribute to assessments of open sea populations, where a variety of models are used to predict future population changes and infer relevant population variables (such as age structure) from the data (Cooper 2006). But these models, if they are involved in RA at all, can only be involved in internal RA. They are models about the same target system that are generated by incorporating extra parameters into a basic model about how the fish population evolves over time. Full-blown external RA is absent here.

Two considerations help to see why it is absent. One is that other, more efficient methods are readily available to measure the relevant properties of the system: the ecological equivalents of Avogadro's number may be slippery or widely dispersed, but they (or at least a sample of them) are usually directly observable. It is not a straightforward exercise to measure the size, mortality, or age structure of a population of ocean fish. But it is considerably less demanding than the measurement exercise that external RA would recommend. External RA might involve using a model of the fish's interaction with another species in the same area, along with current data about that other fish species, to give one determination of the target population's size. Another model, describing the effect of the fish population on a marine plant in the area, along with data about the plant population, could provide a second determination of the fish population's size. This method may, in the end, give a measure of the size of the fish population. But this would come at the cost of constructing two models and collecting two data sets. By contrast, direct measurement of the population size would involve collecting one data set.

A second consideration is that the target systems of fisheries biology may not have the right qualities to support RA. Coincidence RA relies on the presence of a specific causal structure in the target system: two causal processes that are independent except for their shared causal origin. As Levins reminds us, the need for models in population biology is largely due to the complexity of its subject matter. If "complexity" implies "interdependence," it may in general be difficult, in highly complex systems, to find causal processes with the right structure. Consider the method (described in the previous paragraph) of measuring the population size of fish A through fish A's causal interaction with fish population B and marine plant C. It may well be that population B is causally linked to plant C, ruling out the possibility of constructing independent models of the two causal interactions. Not having the experimenter's luxury of isolating and manipulating the target system (fish populations are considerable less portable than collections of molecules) the modeller using external RA must hope that nature puts the target property (the size of the fish A population) in causal contact with other systems that he understands well and that are independent of one another. Fishery biology suggests that this hope is not often realised.

V. Conclusions.

The emphasis by Levins and Weisberg on internal RA lacks motivation, since there is nothing incoherent about external RA. Indeed, consideration of experimental science shows that there exists a fallible but viable form of RA that both confirms robust results and makes use of empirical data.

I have given an account of RA in experimental science, coincidence RA, that shows where empirical data can enter into RA. I have also defended coincidence RA against one objection that purports to show that the empirical data deployed in coincidence RA is not intrinsic to the method. This objection hinges on the notion of independence, the quality that two or more techniques must have in order that a common result of those techniques cannot be attributed to a common feature held between them.

Orzack and Sober's attack on RA also hinges on the notion of independence, and not (*contra* Weisberg) on the incoherence of external RA. Yet Orzack and Sober's chief concern about independence, that "generally speaking, competing models are not independent," ignores the fact that models used in RA do not need to be competing models; that is, they do not need to be models of the same target system. All that is required is that the models are of target systems that are each causally connected to the property or entity that the convergent result of the models describes. The case of RA in experimental science helps to illustrate this requirement, as does the hypothetical example, sketched above, of coincidence RA used in the context of climate change.

The case of fishery biology shows that coincidence RA is not always an appropriate method in model-based science; indeed, the fact that fishery biology uses precise and realistic models of phenomena, and yet does not use coincidence RA, suggests that the method is rarely suitable in model-based science. One lesson from the fishery case is that there may be a link between the reasons for using models in science (the unfeasibility of experiment and the complexity of the phenomena) and the reasons for the failure of RA in the fisheries case (RA is inefficient as a measurement tool, and requires a specific causal structure in the phenomena). A more general lesson is that the fortunes of external RA are sensitive to the resources that are available in the context of its use: in the fishery case, Orzack and Sober's concerns about the non-independence of models seem to hold true, though not quite in the way they describe those concerns. In other fields or disciplines, these concerns may be less relevant. Further investigation into model-based science should shed further light on the conditions that favour external RA. But the first step to evaluating external RA is to acknowledge that it is not incoherent. And the second step is to find out how it performs in ideal conditions, a step that experimental science can motivate and inform.

MICHAEL BYCROFT
mike.bycroft@gmail.com

REFERENCES

Cooper, Andrew B. 2006. *A Guide to Fisheries Stock Assessment: From Data to Recommendation.* University of New Hampshire.

Culp, Sylvia. 1995. Objectivity in Experimental Inquiry: Breaking Data-Technique Circles. *Philosophy of Science*, 62 (3): 438-58.

Eros, Tibor, András Specziár, and Péter Bíró. 2009. Assessing Fish Assemblages in Reed Habitats of a Large Shallow Lake: A comparison between gillnetting and electric fishing. *Fisheries Research* 96(1): 70-76.

Franklin, Allan. 1986. *The Neglect of Experiment.* Cambridge, MA: Cambridge University Press.

Franklin, Allan. 1990. *Experiment: Right or Wrong.* Cambridge, MA: Cambridge University Press.

Franklin, Allan. 2007. Experiment in Physics. *Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/physics-experiment.

Hacking, Ian. 1983. *Representing and Intervening.* Cambridge, MA: Cambridge University Press.

Hudson, Robert G. 1999. Mesosomes: A Study in the Nature of Experimental Reasoning. *Philosophy of Science* 66(2): 289-309.

Levins, Richard. 1966. The Strategy of Model-Building in Population Biology. *American Scientist* 54: 421-31.

Levins, Richard. 1993. A Response to Orzack and Sober: Formal Analysis and the Fluidity of Science. *The Quarterly Review of Biology* 68 (4): 547-55.

Orbach, Peter. 1981. On the Utility of Repeating the Same Experiment. *Australasian Journal of Philosophy* 59: 151-62.

Orzack, Steven and Elliot Sober. 1993. A Critical Assessment of Levins's "The Strategy of Model Building in Population Biology" (1966). *The Quarterly Review of Biology* 68 (4): 533-46.

Perrin, Jean. [1913] 1923. *Atoms*. Translated by D.L. Hammick. Originally published as "Les Atomes" (Paris: Alcan). New York: Van Nostrand.

Rasmussen, Nicolas. 1993. Facts, Artifacts, and Mesosomes: Practicing Epistemology with the Electron Microscope. *Studies in History and Philosophy of Science* 24(2): 221-65.

Rasmussen, Nicolas. 2001. Evolving Scientific Epistemologies and the Artifacts of Empirical Philosophy of Science: A Reply Concerning Mesosomes. *Biology and Philosophy* 16: 629-54.

Salmon, Wesley. 1978. Why ask Why? An Inquiry Concerning Scientific Explanation. *Proceedings and Addresses of the American Philosophical Association* 51: 683-705.

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Staley, Kent. 2004. Robust Evidence and Secure Evidence Claims. *Philosophy of Science* 71: 467-88.

Stegenga, Jacob. Forthcoming. Robustness, Discordance, and Relevance. *Philosophy of Science* 76(5) December 2009.

Weber, Marcel 2005. *Philosophy of Experimental Biology.* Cambridge, MA: Cambridge University Press.

Weisberg, Michael. 2006a. Forty Years of 'The Strategy': Levins on Model-Building and Idealization. *Biology and Philosophy* 21: 623-45.

Weisberg, Michael. 2006b. Robustness Analysis. *Philosophy of Science* 73: 730-42.

Wimsatt, William. 1981. Robustness, Reliability, and Overdetermination in Science. In *Scientific Inquiry and the Social Sciences. A Volume in Honor of Donald T. Campbell,* ed. Marilyn B. Brewer, Barry E. Collins, 124-63. San Francisco: Jossey-Buss.