Model Evaluation: An Adequacy-for-Purpose View

Wendy S. Parker*†

According to an adequacy-for-purpose view, models should be assessed with respect to their adequacy or fitness for particular purposes. Such a view has been advocated by scientists and philosophers alike. Important details, however, have yet to be spelled out. This article attempts to make progress by addressing three key questions: What does it mean for a model to be adequate-for-purpose? What makes a model adequate-for-purpose? How does assessing a model's adequacy-for-purpose differ from assessing its representational accuracy? In addition, responses are given to some objections that might be raised against an adequacy-for-purpose view.

1. Introduction. Twenty-five years ago, Naomi Oreskes, Kristin Shrader-Frechette, and Kenneth Belitz (1994) published a landmark paper in *Science* magazine on the topic of model evaluation. Seeking to combat overconfidence in computer simulation models, they expressed concern about some commonly used terminology, in particular the language of "verification" and "validation." Models cannot be verified or validated in the sense of having the truth of their assumptions established with certainty, they reminded us, nor does impressive past performance by a model guarantee its future performance. Scientific models can be "confirmed" when their output

Received April 2018; revised September 2019.

*To contact the author, please write to: Department of Philosophy, Durham University, 50 Old Elvet, Durham DH1 3HN, United Kingdom; e-mail: wendy.parker@durham.ac.uk.

†For helpful comments and suggestions, the author wishes to thank Donal Khosrowi, Leonard Smith, Adrian Currie, Eric Winsberg, Durham's CHESS research group, and several anonymous referees. This research was supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant 667526 K4U). The content reflects only the author's views, and the ERC is not responsible for any use that may be made of the information it contains.

Philosophy of Science, 87 (July 2020) pp. 457–477. 0031-8248/2020/8703-0004$10.00

is found to fit with observational data, they suggested, but the support provided is "inherently partial." Highly influential, the Oreskes et al. paper has now been cited over 1,800 times, across dozens of fields.[1]

Yet the idea that fit between modeling results and observations "confirms the model" is also problematic. On some accounts, scientific models are structures or objects; they are not the sort of thing that can be true or false and thus are not an appropriate target of confirmation. Even if a model is viewed as a complex hypothesis about the workings of a target system (per Oreskes et al. and many others), it is usually misguided to seek to confirm (or disconfirm or falsify) that hypothesis, since it is usually known from the outset to be false; some of the model's assumptions are known to be highly idealized or simplified, to appeal to fictional entities, and so on (Parker 2010).[2] This is not a merely academic point. If a scientific model is thought to be confirmed when its results are found to fit with observations of a target system, and thereby to accrue credit or to merit increased confidence in some general way, then users can be led to trust other results obtained from the model even when—because of the model's idealized and simplified assumptions—this is not warranted. That is, thinking that one has "confirmed the model" can easily lead to misplaced confidence too.

What should one seek to test—and thereby to confirm or disconfirm or falsify—in the course of model evaluation, if not the model itself? There are a number of reasonable alternatives.[3] One might test whether a model is similar to a target in particular respects and degrees (Giere 1988; Lloyd 2010; Weisberg 2013). Or one might test particular modeling assumptions whose (approximate) truth is still in question, for example, that particular quantities in the target are related according to a certain equation. Or one might test whether a model is adequate or fit for a purpose of interest. This last option is associated with what will here be called an *adequacy-for-purpose* view of model evaluation. On this view, model quality is to be assessed relative to a purpose; model evaluation seeks to learn whether a model is adequate or fit for particular purposes. Such a view can be contrasted with one on which model quality is (just) a matter of how accurately and completely a model represents a target, where the ideal limit is a perfect and complete representation.[4]

1. According to Scopus (https://www.scopus.com), August 2019.

2. This article is concerned with models of natural or social systems/phenomena; much of the analysis also applies in the context of data modeling, as discussed in Bokulich and Parker (2020).

3. These are not mutually exclusive options.

4. See also Teller (2001) on the "Perfect Model Model," Weisberg (2013) on the representational ideal of COMPLETENESS, and Knuuttila (2011, 267) on "the idea that scientific representation should aim for as accurate a representation as possible."

An adequacy-for-purpose view has been advocated by both scientists and philosophers. For example, a US National Research Council report characterizes model evaluation as "the process of deciding whether and when a model is suitable for its intended purpose" (NRC 2007, 3; see also Caswell 1976; Rykiel 1996; Jakeman, Letcher, and Norton 2006; Baumberger, Knutti, and Hadorn 2017). Taper, Staples, and Shepard (2008, 358), discussing modeling in environmental sciences, note "a growing recognition that models should be selected based on their ability to answer questions of interest," rather than on some overall measure of their fit to observational data. Among philosophers, Parker (2010, 287) recommends that model evaluation be understood as "an activity that aims to determine whether a model is adequate for one or more purposes of interest." Currie (2018) adopts a similar view and points to philosophical work on the evaluation of artifacts as a potential resource (e.g., Hilpinen 2011).

An adequacy-for-purpose view is attractive for a number of reasons. It is a natural accompaniment to a prominent way of thinking about scientific models, which emphasizes that models are not just representations but also tools that are selected and used for particular epistemic and practical purposes (e.g., Morrison and Morgan 1999; Giere 2004, 2010; Boon and Knuuttila 2009; Parker 2010; Knuuttila 2011; Currie 2018). In addition, it readily accounts for some prima facie puzzling features of modeling practice, including the fact that modelers sometimes misrepresent a target, even when they could avoid doing so; judicious misrepresentation can promote the achievement of some epistemic and practical aims (Wimsatt 2007, chap. 6; van Fraassen 2008, chap. 1). For example, a modeler might omit from her computer model a (representation of a) particular causal process known to be operating in the target, in order to gain insight into its contribution to the target's behavior. Finally, an adequacy-for-purpose view can help to combat overconfidence in modeling results, just as Oreskes et al. sought to do. This is because, on an adequacy-for-purpose view, evaluators must consider what evidence there is that a model is adequate for the particular purpose at hand; past successes of the model might not be especially relevant.

Despite its attractions, however, an adequacy-for-purpose view of model evaluation has not been the subject of sustained philosophical analysis. A number of basic questions remain at best only partly addressed, including What does it mean for a model to be adequate-for-purpose? What makes a model adequate-for-purpose? How does assessing a model's adequacy-for-purpose differ from assessing its representational accuracy? This article attempts to make progress in answering these questions. In addition, responses are given to some objections that might be raised against an adequacy-for-purpose view. The aim is to further develop, and to defend, a particularly promising view of model evaluation.

The discussion proceeds as follows. Section 2 articulates two notions of adequacy-for-purpose, involving success in an instance of use and reliability in a type of use, respectively, and then suggests a way of understanding the relation between adequacy and fitness. In light of this, section 3 draws out a key insight: for a model to be adequate-for-purpose, it must stand in a suitable relationship not just with a representational target but with a target, user, methodology, and background circumstances jointly. Section 4 outlines some basic strategies for assessing adequacy-for-purpose and identifies several ways in which assessing adequacy-for-purpose differs from assessing (just) a model's representational accuracy. Section 5 responds to a number of potential objections to an adequacy-for-purpose view. Finally, section 6 offers some concluding remarks.

2. Characterizing Adequacy-for-Purpose. Advocates of an adequacy-for-purpose view have said little about what it means for a model to be adequate-for-purpose. For instance, Parker (2011, 584) says only: "An adequate model is one that is sufficient for the purposes of interest not just as a matter of accident (e.g., a one-off accurate prediction) but because the model has properties that make it suitable for those purposes." For Currie (2018, 775), an adequate model is one that has properties that "promote the kind of model output which is desired."[5] After some brief remarks on purposes, this section articulates two notions of adequacy-for-purpose and suggests a way of understanding the relation between adequacy- and fitness-for-purpose.

2.1. Purposes. As understood here, a purpose is a goal. Very often, the purposes for which models are used are epistemic—they are used to predict something, to explain something, to teach something, and so on. In some cases, purposes have built into them constraints on how an outcome should be achieved. For example, the goal might be not just to correctly predict whether X will occur but to do so for roughly the right reasons, that is, because one has taken account of the main factors that actually determine whether X will occur.

Sometimes, the stated purposes for which models are used are practical. The goal might be to increase the profits of a firm or to protect a population from some natural hazard. Even in these cases, however, the intended contribution of the model is often epistemic: it is expected that the model's serving one or more epistemic purposes will, in the context of a more extended activity, facilitate the achievement of the practical purpose. In a

5. See also Currie's three criteria for evaluating "the success of an artifact" (2018, 774), which seem broader, since they include such things as whether the artifact/model matches the design that its creator had for it.

study (Haasnoot et al. 2014) that will be revisited at various points below, for example, the ultimate goal is to ensure efficient flood protection and a fresh water supply in the Netherlands' Rhine river delta (a practical purpose), but the model's role is to predict, in a coarse-grained way, the water-related consequences of various policy choices for the region (an epistemic purpose).

As the Haasnoot et al. example illustrates, the purposes for which models are constructed and used are often rather circumscribed and local (Alexandrova 2010). Further examples of such purposes could be (1) to accurately predict the development of hurricanes in the Atlantic Ocean, (2) to explain how a drug inhibits neural activity in a particular part of the brain, (3) to identify the source of an unusual contaminant in a particular river, (4) to increase Illinois high school students' knowledge of world geography, and (5) to explore the implications of a new theory of adolescent group dynamics. These examples also illustrate that, while some purposes can be achieved to a greater or lesser extent (e.g., 1), others can only be achieved or not (e.g., 3).

Purposes often are stated in ways that permit multiple interpretations. What counts as successfully explaining how a drug inhibits neural activity in a particular part of the brain, for example, will depend on one's views on explanation. In such cases, disambiguation will be an important first step in model evaluation: in order to judge a model's adequacy or fitness for a purpose, the evaluator needs to understand what the purpose is and what would count as successfully achieving it. This might require significant unpacking and reflection.

2.2. Adequacy-for-Purpose. To ask whether a model is adequate for a purpose is to ask something like: Can the model be used to do the job? But this also can be interpreted in various ways, suggesting that a number of different conceptions of adequacy-for-purpose could be articulated. For instance, there could be a very weak notion of adequacy-for-purpose, according to which a model is adequate for a purpose P if there is some possible way of using the model, such that someone could at least sometimes achieve P. Below, however, two stronger conceptions are articulated, which seem of greater interest in scientific practice.[6]

First, there is a notion of adequacy associated with *success in a particular instance of use*:

> **ADEQUACY$_1$:** A tool M is ADEQUATE$_1$-FOR-P if and only if using M in instance I results in the achievement of purpose P.

6. These conceptions are meant to be applicable not just when evaluating scientific models but when evaluating other tools as well. Even in modeling contexts, the tool of interest is sometimes a set of models (e.g., used for the purpose of quantifying uncertainty) or a larger system that incorporates a model (e.g., a forecasting system).

Any instance of use of a tool involves a user U, a way W in which the tool is used, and background circumstances B in which the use occurs; it takes place somewhere and at some time. Suppose that a climate scientist U uses a climate model M in an attribution study (following methodology W and in circumstances B) with the aim of P: discerning whether most of the global warming observed since 1950 is due to rising anthropogenic greenhouse gas concentrations (see, e.g., IPCC 2013, chap. 10). Suppose that she reaches an affirmative conclusion in light of the study results and that this conclusion is correct, that is, that anthropogenic greenhouse gases did in fact cause most of the warming. Then, the model is ADEQUATE$_{I}$-FOR-P. Even if, using M in a similar way, the climate scientist would often fail to reach correct conclusions about the contribution of anthropogenic greenhouse gases to other observed changes in climate-related variables, in this instance she succeeded.[7]

A second notion of adequacy-for-purpose is associated with *reliability in a type of use*:

> **ADEQUACY$_{C}$:** A tool M is ADEQUATE$_{C}$-FOR-P if and only if, in C-type instances of use of M, purpose P is very likely to be achieved.[8]

A type of use—also referred to here as a *context of use*—can be specified in terms of one or more users or types of user U, one or more methodologies or types of methodology W, and some range of circumstances of use B. Consider, for example, a computer model M designed to simulate the storm surge (local rise in sea level) that occurs along the US coast when a hurricane is nearby. The model developers would like M to be ADEQUATE$_{C}$-FOR-P: predicting whether the surge at particular coastal locations will exceed prespecified hazardous levels during a 48-hour forecast period.[9] The context of interest (C) involves operational forecasters U at the US National Hurricane Center, who follow a particular methodology W for generating predictions with M and in usual background circumstances B (without power cuts to their computers, in a climate like today's, etc.). If M is ADEQUATE$_{C}$-FOR-P, then, whenever

7. The wording is important: while it is a tool that is ADEQUATE-FOR-P (or not), it is always the user, not the tool, who achieves or fails to achieve purpose P.

8. Just how likely the achievement of P needs to be can vary somewhat from case to case, depending on, e.g., the importance of achieving P. There are various ways to understand this probability: as the frequency of success in achieving P in a hypothetical, long-run series of repeated trials of type C, as a propensity, etc. No stand is taken on this here.

9. In reality, such forecasts might be probabilistic, and the goal might be more modest, e.g., to improve on current forecast performance. However, because evaluating the accuracy and improvement of probabilistic forecasts is a complicated and contested matter, the example here takes a simpler form.

the forecasters use M in way W and in B-type circumstances, they are very likely to predict correctly whether the surge will exceed the hazardous thresholds at those locations.[10]

As the storm surge example suggests, scientists often will be interested in whether a model is ADEQUATE$_c$ in a context in which it is actually used; we might call this its *adequacy-in-practice* for P. Sometimes, however, they might be interested in whether a model is ADEQUATE$_c$ in other contexts of use—in contexts that they aspire to (e.g., once they have the resources to implement an improved methodology) or even in ideal contexts (e.g., in which highly competent and conscientious users follow an ideal methodology, in highly favorable circumstances; we might call this its *adequacy-in-principle* for P). ADEQUACY$_c$ as articulated above can accommodate these different cases. Likewise, evaluators could consider in which contexts of use a model is ADEQUATE$_c$ for a given purpose or for what range of purposes there is a practically accessible context of use in which the model is ADEQUATE$_c$, and so on.

Can a model be ADEQUATE$_c$ as a matter of luck or coincidence? It seems unlikely, assuming that contexts of use are not specified so narrowly that they encompass just a few possible instances of use. It seems more plausible that, if the use of a model M across a range of instances would almost always result in the achievement of P, it is because M has properties that facilitate in some systematic way(s) the achievement of P in that context (reminiscent of the suggestions of Parker and Currie). If the storm surge model is ADEQUATE$_c$ for a predictive purpose P, for example, it might be in part because it represents accurately enough the physical processes that actually control storm surge in those locations.

2.3. Fitness-for-Purpose. The concept of fitness-for-purpose is invoked in discussions of model evaluation at least as often as that of adequacy-for-purpose, although it too usually goes unanalyzed. Here, fitness-for-purpose is analyzed in terms of adequacy-for-purpose. Let x stand for either the instance (I) or type (C) variety of adequacy-for-purpose articulated above:

FIT$_x$: A tool M is FIT$_x$-FOR-P if and only if it is ADEQUATE$_x$-FOR-P.

Unlike adequacy, however, fitness is a concept that admits of degrees. This is useful when evaluating models for purposes that can be achieved to a greater or lesser extent. Consider an example given earlier, P: increasing Illinois high school students' knowledge of world geography. For this purpose, even a small increase in knowledge will count as a success, but larger

10. Even if the storm surge model is never actually used in a C-type instance, it might still be ADEQUATE$_c$-FOR-P. ADEQUACY$_c$ is a modal notion; it is a matter of what would be the case, if C-type instances of use were to occur.

increases are more desirable. We can think of such purposes as having a more complex structure, consisting of a rank-ordered set of achievements, $\{P_{min}, ..., P_{max}\}$, where the first (and lowest-rank) member of the set, P_{min}, corresponds to achieving P to the minimally acceptable extent and the last (and highest-rank) member of the set, P_{max}, corresponds to achieving P to the maximally desired extent. Then:

> **FITNESS$_X$:** A tool M's FITNESS$_X$-FOR-P is greater to the extent that M is ADEQUATE$_X$-FOR-P for a higher-ranking member of $P = \{P_{min}, ..., P_{max}\}$.

Thus, if the increase in Illinois high school students' knowledge of world geography that would result from using a model M_1 in a context of use C would almost always be at least P_m, while the increase resulting from the use of M_2 would almost always be at least P_n, and if $P_m > P_n$, then M_1's FITNESS$_C$-FOR-P is greater than M_2's. Claims of relative fitness can be made even if the context of use of M_2 differs from that of M_1, as long as the ranking $\{P_{min}, ..., P_{max}\}$ is the same.[11]

3. What Makes a Model Adequate-for-Purpose? We now are better positioned to address another fundamental question: What makes a model adequate-for-purpose? That is, in virtue of what is a model adequate for a purpose P? Clearly, the specific features that make a model adequate-for-purpose will vary from case to case. Nevertheless, the conceptions of adequacy-for-purpose articulated in the last section point to an important general insight about what makes a model adequate-for-purpose.

In particular, for a model to be adequate-for-purpose, it must stand in a suitable relationship not just with a representational target T but with a target T, user U, methodology W, circumstances B, and goal P jointly. The model must have features, including but not limited to how it represents target T, such that user U, using the model in way W in circumstances B achieves (or is very likely to achieve) purpose P.[12] We can also think of T, U, W, B, and P as dimensions of a *problem space*, constituted by a goal (P) and a set of constraints (T, U, W, B) on how that goal should be achieved;

11. Given that fitness-for-purpose is analyzed in terms of adequacy-for-purpose, most of the remainder of the discussion will be framed just in terms of adequacy-for-purpose, for simplicity. Likewise, where a point applies to both ADEQUACY$_I$ and ADEQUACY$_C$, I will just speak of "adequacy-for-purpose."

12. It is interesting to consider how the two notions of adequacy outlined above relate to conceptions of singular and general causation and whether the use of a model, or even the model's relation to its target, might be understood as INUS conditions (an insufficient but necessary part of an unnecessary but sufficient set of conditions) for the achievement of the purpose. Space limitations prevent exploring this here.

TABLE 1. Examples of Inadequacy due to Lack of Fit with Users, Methodology, or Background Conditions

Type of Purpose	Type of Model	Reason for Inadequacy
Pedagogical	Physical	M is very sensitive; vibration in environment disrupts M's functioning (B)
Explanatory	Mathematical	Although M's equations are very accurate, they are so complex that explanatory information is not salient to users (U)
Predictive	Computational	An idealization in M amplifies, rather than compensates for, common errors in initial conditions estimated from data (W)

model M needs to be a "solution" within that problem space.[13] A great deal of philosophical discussion of scientific modeling has focused on when models represent, or successfully represent, their targets—pointing to relationships of isomorphism, partial isomorphism, homomorphism, similarity, exemplification, and more (see Frigg and Nguyen [2017] for a critical review). Yet the model-target relationship is only part of the story when it comes to achieving scientific aims; the model's features must also align with those of U, W, and B. As table 1 illustrates, a model can fail to be adequate-for-purpose for reasons that relate to these other dimensions of the problem space as well.

Note that this is not just the simple (but correct) point that pragmatic features of a model, such as how long it takes to run on a computer, can matter in addition to a model's representational features. There can be interactive effects, in the sense that how a model should represent a target T, if it is to be adequate-for-purpose, can depend on features of U, W, and B.[14] This is illustrated in the last two entries of table 1, where a model fails to be adequate-for-purpose because its representational features fail to align with features of the user U and the broader methodology W, respectively.[15] In the latter,

13. Related concepts can be found in cognitive science and in engineering design (see, e.g., Goel and Pirolli 1992 and references therein). In a similar vein, Mayo (2018, 297) speaks of statistical models of data being "adequate for a problem" and statistical hypotheses being "conjectured solutions" to a problem of statistical inference. Target T is included among the dimensions of the problem space here because P is to be achieved using some model of T; for problems that are to be solved using nonrepresentational tools (e.g., hammers), the problem space would not include any T.

14. Various philosophers and scientists note the importance of features other than how a model represents a target. For some recent examples, see Jakeman et al. (2006), Mäki (2009), Oberkampf and Roy (2010), Elliott and McKaughan (2014), Bolinska (2016), and Baumberger et al. (2017). The current article provides a framework for thinking more broadly and systematically about this—in terms of constraints imposed by features of U, W, and B—including constraints on the representational features of models.

15. As the example involving explanation illustrates, increasing the fidelity or realism of a model does not necessarily increase its fitness-for-purpose, nor does it necessarily

a computer model incorporates an idealization that amplifies errors that occur in the initial conditions from which the model is run, resulting in predictions that do not display sufficient accuracy; the same model might be used successfully for that predictive purpose in a problem space that includes a different methodology for estimating initial conditions.

The Haasnoot et al. (2014) study mentioned earlier provides a real-world illustration of the relevance of more than how a model represents a target. The authors evaluate the adequacy of a computational model for the purpose of (P) screening and ranking different water policy pathways for the Rhine river delta through the end of the twenty-first century, a period in which climate is expected to change by an uncertain amount. They emphasize that, for their purposes, the model not only needs to give sufficiently accurate results for a range of variables related to floods, droughts, and their impacts but also needs to be computationally efficient, so that numerous courses of action can be explored in a limited time period (a 5-year decision-scoping program). After providing some evidence that their model does meet these basic requirements, Haasnoot et al. also praise the model's transparency and adaptability (111), citing these as features that will facilitate exploring policy options in an interactive way with stakeholders. Their evaluation thus considers not just how the model relates to a representational target but also whether it stands in a suitable relationship with users and with the computational methodology to be employed.

It is important to recognize, however, that although T, U, W, and B constrain the features that a model should have if it is to be adequate-for-purpose, they rarely uniquely determine them. Often there will be many possible ways to construct an adequate model. A predictive purpose, for instance, might be achieved by keeping errors in all relevant variables in a mathematical model small or by allowing larger errors but ensuring that they mostly cancel out, and so on. In fact, the features that make one model adequate-for-purpose can be very different from those that make another model adequate for that same purpose. Suppose two scientists want to learn whether a new drug will be more effective than an alternative in alleviating particular symptoms of a disease in a given population. One scientist investigates this via an experimental study involving animal models, while the other uses a computational model that simulates the molecular chemistry of the drug, that is, how it binds and interacts with particular molecules that are related to the disease. In both cases, the models might be adequate-for-purpose. Yet the features that make the animal model adequate—which include biological traits of the animals—are obviously quite different from the features that make the computational model adequate.

---

warrant an increase in one's confidence that the model is adequate-for-purpose. See also Levins's (1966) claims about trade-offs among precision, generality, and realism in modeling.

Thus, when conceptualized in terms of adequacy-for-purpose, model quality is even more context dependent than is usually emphasized. It is not just that the features that will make for an adequate model depend on the purpose (Teller 2001); even once the purpose is specified, the features that will facilitate the achievement of that purpose can vary, depending on the user, methodology, the circumstances of use, and perhaps even the type of model being employed (e.g., physical vs. mathematical). This does not necessarily mean, however, that nothing useful can be said (about what makes a model adequate-for-purpose) at a level more general than that of the individual case. It is an open question whether some midlevel theory can be developed that, for different types of purpose, methodology, user, or model, offers some important insights about what makes a model adequate.[16]

4. Assessing Adequacy-for-Purpose. A third question of interest is: How does assessing a model's adequacy-for-purpose differ from evaluating (just) its representational accuracy? In practice, model evaluation—however it is conceptualized—often begins informally as a model is being developed and only later becomes a more formal activity. At the formal evaluation stage especially, it can involve a host of subtle issues that depend on fine details of the case at hand, for example, what sort of calibration or tuning of the model was performed, differences between the spatial or temporal resolution of modeling results and observations to which they are compared, specification of parameters in statistical tests, and so on (see, e.g., NRC 2007, chap. 4; Oberkampf and Roy 2010). It is impossible to do justice to these complexities here. Instead, this section provides an overview of some basic strategies that can be employed in the assessment of adequacy-for-purpose (sec. 4.1), making it easier to see several significant ways in which assessments of adequacy-for-purpose can differ from assessments of representational accuracy (sec. 4.2).[17]

4.1. Basic Strategies. In general terms, to assess adequacy-for-purpose is to consider what reasons there are for thinking that a model is (or is not) adequate for a purpose of interest. Both information about how a model is constructed—that is, about its ingredients or components and how they are arranged—and information about a model's performance can be relevant (see also Baumberger et al. 2017). Ultimately, the evaluator seeks to determine the appropriate level of confidence to have in (or whether to accept or

16. Weisberg's (2013) analysis of the kinds of model-target similarities that are important for different types of modeling goal could be interpreted as a first step in this direction (see Parker 2015; Jacquart 2016).

17. Note that model evaluation need not always be stringent or formal. More careful assessment is desirable when erroneously accepting that a model is/is not adequate for a purpose of interest can be foreseen to have significant negative consequences.

reject) a hypothesis about the model's adequacy-for-purpose. For example, the hypothesis of interest might take the form, H: model M is ADEQUATE$_c$ for purpose P.

Construction. When considering the construction of a model, the evaluative task is akin to assessing the design of a tool with respect to an envisioned problem space; the question is whether the model/tool has properties that will facilitate the achievement of the purpose in a context or instance of use (see also Currie 2018). It is easy to see how information about model construction can be relevant. To give an extreme example, if an evaluator learns that a climate model has a slab ocean—one that does not represent any internal ocean dynamics—she can be very confident that the model is not ADEQUATE$_c$ for (P) gaining insight into how ocean dynamics will change as atmospheric greenhouse gas concentrations rise. By contrast, if an evaluator learns that an animal model (e.g., a mouse species) has a particular set of biochemical pathways in common with humans (i.e., a key similarity with T) and is easy to house and care for in a local laboratory (i.e., aligns with U, W, B), this might increase her confidence that the model will be ADEQUATE$_i$ for (P) learning whether a new drug is efficacious in treating a particular disease in humans. This can also be expressed in terms of *confirmation*: the information about the animal model confirms (i.e., provides some support for) the hypothesis that the model is ADEQUATE$_i$ for P.[18] Likewise, the information about the climate model strongly disconfirms the hypothesis that the model is ADEQUATE$_c$ for P; indeed, it seems to provide grounds for rejecting the hypothesis altogether.

Performance of Model Components. Evidence for or against a model's adequacy-for-purpose can also be obtained by examining the performance of model components with respect to observational data or other benchmarks. In some engineering contexts, for instance, component-level testing of computational models is carried out in a rigorous way, with performance requirements specified in advance in light of the purpose for which the model as a whole is to be used (see, e.g., Thacker et al. 2004; Oberkampf and Roy 2010). It might be specified, for example, that results for variable X in model component Y should not differ from observed values of X by more than 5%. Note, however, that whether the performance of a model component supports the hypothesis that the model is adequate-for-purpose can depend on how other components of the model are expected to perform

18. The nature and strength of the support depend on the view of confirmation that is adopted. The error statistician takes a different approach, asking whether the hypothesis that the model is adequate-for-purpose has passed a severe test with some data e (see, e.g., Mayo 1996, 2018).

too. Large errors in results from one component might not be evidence against the model's ADEQUACY$_c$ for some purpose P, for instance, if those errors are expected to be systematically compensated for elsewhere in the model.

Direct Testing of Adequacy. Sometimes it is possible to directly test a hypothesis about a model's adequacy-for-purpose. When the hypothesis concerns a model's ADEQUACY$_c$, direct testing involves examining the model's performance in a sample of C-type instances of use. For example, to directly test the hypothesis (H) that model M is ADEQUATE$_c$ for (P) predicting with at least accuracy *A*, the power output for a particular set of wind turbines, an evaluator might check whether, in a sample of C-type instances of use, predictions obtained using M almost always display at least accuracy *A*; if so, this can provide some support for H. The strength of the support will depend on the how large the sample is and whether it can be considered a random sample of C-type instances of use or a sample in which it would be particularly easy or difficult to achieve *A*. Likewise, sometimes an evaluator can directly test a model's ADEQUACY$_i$ for a purpose P by simply waiting to see whether P is achieved in that instance. In this way, an evaluator might obtain very strong evidence regarding the ADEQUACY$_i$ of model M for (P) predicting tomorrow's wind power output with at least accuracy *A*.

Often, however, direct testing of a model's adequacy-for-purpose is undesirable or even impossible. Scientists do not want to wait until a breach occurs at a local nuclear reactor to learn whether their computer model is adequate for simulating, with a specified level of accuracy, how radioactive materials would disperse in the case of such a breach. And direct testing is impossible when models are used—as they frequently are—to investigate counterfactual situations that will never be realized, for example, when climate models are used to project changes in climate under multiple future greenhouse gas emission scenarios; at most one of those scenarios will be realized.

Indirect Testing of Adequacy. In such cases, an evaluator sometimes can indirectly test a model's adequacy-for-purpose by examining its performance in other instances or contexts of use. Scientists assessing the adequacy of their model for simulating the dispersal of radioactive materials in the vicinity of a local reactor, for example, might examine how well their model can simulate the dispersal of other materials in other locations, for which some observational data are available. When such indirect tests are performed, however, evaluators need to take account of differences between the test situations and the instances or contexts of use that are ultimately of interest. In the case of the dispersal simulations, for example, perhaps the available test situations involve relatively simple topography, while the topography near

the local reactor is more complex. In that case, the model's performing well in the test situations might provide only weak support for (or weak confirmation of) its adequacy-for-purpose. Unfortunately, it is sometimes quite challenging to determine to what extent a model's performance in one instance or context of use constitutes evidence for or against its adequacy-for-purpose in another (Parker 2010, 2011).

Synthesis. In practice, assessments of adequacy-for-purpose usually involve several of the basic strategies just outlined (and sometimes others too). In the Haasnoot et al. (2014) study, the assessment considered not only the model's design—whether it included relevant variables and parameters in its representation of the water system of the Rhine delta (the target) as well as its computational efficiency, transparency, and adaptability—but also the performance of the model and its components when compared to observational data and to results from more complex models. For example, the evaluators checked whether, in simulations for past years for which observations are available, the model's water distribution module (a component) correctly indicated whether water levels in a key set of lakes were above or below thresholds where significant damage begins to occur. Direct testing of the model's adequacy-for-purpose was not possible, since the model was being used to make long-term predictions of conditions under a number of different climate change scenarios, at most one of which will eventually be realized. Instead, evaluators performed a series of indirect tests, checking whether the model responded to interventions in ways that experts and stakeholders considered plausible, for example, whether changing features of dikes in the model resulted in plausible reductions in flood damages.

Such findings need to be aggregated to reach some conclusion about a model's adequacy for a purpose of interest. In practice, this aggregation step is often left implicit. Haasnoot et al. (2014, 112), for example, conclude that their model is adequate-for-purpose after reaching affirmative answers to a series of questions about the model's design and performance, but they do not explain why affirmative answers to this set of questions are together taken to be sufficient. Similarly, Baumberger et al. (2017) articulate a set of general considerations that can be appealed to when arguing that a model is (or is not) adequate for a predictive purpose, but they say little about how to weigh up or combine those considerations; their remarks are suggestive of an informal Bayesian perspective, where confidence in a model's adequacy-for-purpose is increased (or decreased) in light of information about a model's construction or performance but without formal, quantitative Bayesian updating (see also Schmidt and Sherwood 2015).[19] Further work is needed to explore how philosophical perspectives on evidence, such as Bayesian

19. Baumberger et al. (2017, 15) also speak of obtaining "premises for a nondeductive argument for the claim that a model is adequate."

and error-statistical perspectives, could be helpful in conceptualizing and guiding the aggregation of evidence in model evaluation.

4.2. How Assessing Adequacy-for-Purpose Is Different. The strategies just discussed point to a number of ways in which the assessment of adequacy-for-purpose differs from assessment that focuses (just) on a model's representational accuracy. Four such differences are identified here; this is not an exhaustive catalog.

First, and most obviously, while evaluating a model's representational accuracy requires that the evaluator consider how a model fits a target, evaluating a model's adequacy-for-purpose requires that the evaluator consider whether a model stands in a suitable relationship with a problem space, which encompasses a target T, (type of) user U, (type of) methodology W, (type of) circumstances B, and goal P. This is a different evaluative task, involving a *broader range of considerations*. The additional factors to be considered—such as the properties of the model user—can be another source of uncertainty in the evaluative process. When considering model-target fit, the adequacy-for-purpose evaluator will focus on aspects (and degrees of fit) that are considered most relevant for achieving the purpose of interest, whereas the evaluator of representational accuracy might well employ general or overall measures of model-target fit.

Second, evaluating a model's adequacy-for-purpose involves a kind of *holism* that is absent when evaluating a model's representational accuracy. For any given aspect of a target that is represented in a model, one can ask how accurately the model represents that aspect; in principle, information about a model's representational accuracy can be accumulated one aspect at a time. By contrast, when assessing a model's adequacy-for-purpose, aspects of the model often cannot be assessed independently (see also Lenhard and Winsberg [2010] on "fuzzy modularity"). As noted in section 4.1, whether errors in results from one component of a model speak against the model's adequacy-for-purpose can depend on whether those errors are compensated for elsewhere in the model. It can even depend on the broader methodology in which the model is embedded, since that methodology might include corrective steps, for example, as when output from weather-forecasting models are postprocessed to correct for known biases.[20]

20. Rice (2019, 196) contends that many highly idealized models "should be characterized as holistically distorted representations of their target system(s) that are greater than the sum of their accurate and inaccurate parts," but it is unclear whether he means to deny that information about a model's representational accuracy can be accumulated one aspect at a time. He seems more concerned to show that "the explanations and understanding provided by scientific models are typically the result of a rich and complicated mixture of various modeling assumptions whose contributions cannot be studied in isolation" (196), which accords well with the current analysis.

A third and closely related difference stems from the fact that there is often more than one way to construct a model that is adequate for a purpose of interest. It is clear how a model should perform with respect to (accurate) observations of a target if the model is an accurate representation of that target: there should be a very good fit. This is so for any aspect of the target represented in the model. By contrast, when testing whether a model is adequate for a purpose, the expected fit between modeling results and accurate observations of the target is often *significantly underdetermined* unless a hypothesis is also made about how the model manages to be adequate for that purpose. If the purpose is a predictive one, for instance, the model might be adequate because errors in contributing variables are all small or because relatively large errors in contributing variables systematically compensate for one another, and so on. Without assuming one of these to be the case, it will be unclear how well the model's results should fit observations of a given contributing variable, if the model is adequate-for-purpose (Parker 2009). Putting a positive spin on this, when the evaluator compares the model's results for the contributing variables to observations, she might obtain evidence not only that the model is adequate-for-purpose but also about how it manages to be adequate.

Finally, unlike when assessing whether a model is an accurate representation of a target, *differences in the testing context can matter* when assessing adequacy-for-purpose. Results from a model that is an accurate representation are expected to fit well with (accurate) observations of the target, whenever and wherever the model is tested. By contrast, the performance that an evaluator should expect from a model if it is adequate-for-purpose can vary tremendously with the test situation, because of differences in users, methodologies, circumstances, or purposes. A weather-forecasting model that is $\text{ADEQUATE}_{\text{C}}$ for a predictive purpose P might give forecasts that fit well with observations in C-type instances of use but might give forecasts that fit less well with those same observations in a slightly different context of use C', which involves a small change in the methodology W that is employed, such as a reduction in the range of information used to estimate initial conditions for the model. As noted in section 4.1, determining how a model's performance should be expected to differ across instances or contexts of use if it is adequate-for-purpose can be challenging.

5. Worries and Replies. There are a number of worries that might be had about an adequacy-for-purpose view. Those addressed here are that an adequacy-for-purpose view is so demanding as to be useless, that it is too narrow, that it wrongly blames models for failures that are not their fault, and that it obscures the importance of representational accuracy.

Worry.—When assessing adequacy-for-purpose, the evaluator must consider how a model relates to a target, user, methodology, circumstances, and

goal jointly. This is often difficult, for reasons discussed above. The worry arises that, despite the best efforts of evaluators, it will very often remain unclear whether a model is adequate for purposes of interest. This seems to be the case, for example, when evaluating the adequacy of climate models for predicting long-term changes in climate to within specified margins of error (Parker 2009; Katzav 2014). This in turn might lead one to doubt the usefulness of an adequacy-for-purpose view itself. This is the position taken by Katzav (2014) for the case of climate modeling. He proposes that, rather than adopting an adequacy-for-purpose view, the assessment of climate models should focus on whether results of interest from the models are "real possibilities" (236). A real possibility is, very roughly, something that we do not have good reason to think could not be the case (see Katzav 2014 for details).

Reply.—The approach that Katzav recommends seems easily recast in adequacy-for-purpose terms: climate models should be evaluated with respect to their adequacy for the purpose of revealing real possibilities about (particular aspects of) future climate change, rather than for the purpose of giving predictions to within specified margins of error. This is because we are in a better position to reach confident conclusions about the former than about the latter (Parker 2009). In fact, the climate modeling example points to a general strategy available to adequacy-for-purpose evaluators: if it is unclear whether a model is adequate for one purpose of interest, consider whether there is evidence that the model is adequate for some related, easier-to-achieve purpose that is also of interest (e.g., revealing something possible or plausible, rather than showing what will happen). In some cases, of course, evaluators may remain highly uncertain whether a model is adequate for any purposes that are currently of significant interest. But this is not a shortcoming of an adequacy-for-purpose view; it is just an unfortunate epistemic reality.

Worry.—The adequacy-for-purpose view is too narrow to serve as a general account of model evaluation; it may be apt for applied science contexts, but not for science in general, since sometimes the aim of modeling is simply to develop an accurate representation of aspects of a target, not to predict, explain, teach, and so on.

Reply.—This assumes that accurate representation—and presumably also the closely related goal of accurate description—cannot be purposes of interest for which the adequacy of a model is evaluated. It is unclear what would justify this assumption, and rejecting it dissolves the worry. It is also worth noting here that an adequacy-for-purpose view does not require that every model evaluation activity be explicitly linked to a particular purpose. An evaluator might first spend time learning how accurately a model represents a target in various respects—or, to return again to the discussion of section 1, learning how similar the model is to the target in various respects or

whether particular modeling assumptions are true of the target—in order to put herself in a better position to reach conclusions later about the range of purposes for which the model is adequate. An adequacy-for-purpose view simply says that, ultimately, the aim of model evaluation is to learn about a model's adequacy or fitness for purposes of interest.

Worry.—The adequacy-for-purpose view wrongly blames models for failures that are not their fault. Consider a model M that is a very-high-fidelity representation of a target. On the analysis given here, M might be inadequate for P in some instances or contexts of use for reasons that have to do with the methodology, user, or circumstances—because the methodology employed introduces errors into the final results produced, because the user's cognitive abilities are too limited to appreciate the explanatory information that M could provide, and so on. Surely, the worry goes, there is something wrong with a view that declares high-fidelity model M to be inadequate in such cases, when the real problem seems to lie elsewhere.

Reply.—A first reply is that an adequacy-for-purpose view would not declare M to be inadequate tout court or even inadequate for P; it would merely declare M to be inadequate for P in those instances and contexts of use. Such a high-fidelity model no doubt would be adequate for P in many other possible instances and contexts of use (see sec. 2.2). A second reply points to scientific practice. Sometimes scientists want to know whether a model does the job in the instance or context of use in which they actually employ it. Even if, say, their methodology is imperfect in various ways, it might be the best that they can currently manage, and they want to know whether they nevertheless will (or can reasonably expect to) succeed in achieving their goal when using their model in accordance with that methodology. In such cases, it seems entirely appropriate that positive and negative evaluations should be pegged to whether the model facilitates achievement of the goal in that instance or context of use. Finally, there is nothing in the adequacy-for-purpose view that says that, when a model is inadequate in a given instance or context of use, the best way forward is to "blame the model." On the contrary, one might have good reason to think that adjusting other aspects of the problem space—improving the methodology or taking steps to avoid background interfering factors—would put one in a better position to achieve the purpose (and others) in the future.

Worry.—An adequacy-for-purpose view obscures the value of accurate representation. It seems to imply that features like a model's computational complexity or whether a model is easy for a user to manipulate are just as important as how the model represents the target. But, one might think, it is how a model represents its target that is really most important.[21] In the long run, as various practical constraints are overcome (e.g., as computing power

21. Thanks to an anonymous referee for suggesting an objection along these lines.

increases), how a model represents a target will be all that matters, and, moreover, high-fidelity models will be most desirable—they will be adequate for a broad range of scientific purposes.

Reply.—The defender of an adequacy-for-purpose view can readily agree that, for many scientific purposes, high-fidelity models are desirable, not just in the long run but today. This does not mean that high-fidelity models are always the best tool for the job. And while some practical constraints will be relaxed as technology develops, other reasons for sometimes preferring lower-fidelity models—such as human cognitive limitations or ease of use—may well persist even in the long term. In any case, the adequacy-for-purpose view is meant to be relevant and useful for today's science, not for some imagined, future science. In today's science, a model's having the wrong "practical" features can stop scientists from achieving their goals just as much as the model's standing in the wrong relation to the target can; in that limited sense at least, they are equally important. It is a strength of the adequacy-for-purpose view that it does not render such "practical" considerations a mere afterthought in model evaluation.

6. Conclusion. This article aimed to flesh out and defend an adequacy-for-purpose view of model evaluation. Three main questions were addressed. First, what does it mean for a model to be adequate-for-purpose? Two varieties of adequacy were introduced, one concerned with success in a particular instance of use (ADEQUACY_i) and another concerned with reliability in a type or context of use (ADEQUACY_c). Using these basic notions, we can define further varieties (e.g., *adequacy-in-principle*) and can also articulate notions of fitness-for-purpose. Second, what makes a model adequate-for-purpose? The key insight here was that, while the specific features that make a model adequate-for-purpose vary from case to case, in general terms what is required is that the model stands in a suitable relationship with a target, (type of) user, (type of) methodology, (type of) circumstances, and purpose jointly. Put differently, the model must constitute a "solution" in a kind of problem space. Third, how does assessment of a model's adequacy-for-purpose differ from assessment of (just) a model's representational accuracy? We saw that it involves a broader range of considerations and can involve special challenges related to holism, underdetermination, and context dependence in testing.

Clearly there is room for further development of an adequacy-for-purpose view. One open question is whether there are varieties of adequacy-for-purpose, beyond those articulated here, that are particularly important in scientific practice. In addition, it remains to be seen whether a useful midlevel theory—considering different types of purpose, user, methodology, circumstances, and model—can be developed to shed further light on the sorts of features that make a model adequate in different cases. Finally, it would be

helpful to have a number of detailed case studies exploring how the evaluation of adequacy-for-purpose does or could proceed in practice in different scientific contexts, recognizing both formal and informal evaluation practices.

REFERENCES

Alexandrova, Anna. 2010. "Adequacy-for-Purpose: The Best Deal a Model Can Get." *Modern Schoolman: A Quarterly Journal of Philosophy* 87 (3–4): 295–301.

Baumberger, Christoph, Reto Knutti, and Gertrude Hirsh Hadorn. 2017. "Building Confidence in Climate Model Projections: An Analysis of Inferences from Fit." *WIREs Climate Change* 8:e454.

Bokulich, Alisa, and Wendy S. Parker. 2020. "Data Models, Representation, and Adequacy-for-Purpose." Unpublished manuscript, Boston University.

Bolinska, Agnes. 2016. "Successful Visual Epistemic Representation." *Studies in History and Philosophy of Science* A 56:153–60.

Boon, Mieke, and Tarja Knuuttila. 2009. "Models as Epistemic Tools in Engineering Sciences: A Pragmatic Approach." In *Handbook of the Philosophy of Science*, vol. 9, *Philosophy of Technology and Engineering Sciences*, ed. Antoni Meijers, 687–720. Amsterdam: Elsevier.

Caswell, Hal. 1976. "The Validation Problem." In *Systems Analysis and Simulation in Ecology*, vol. 4, ed. Bernard C. Patten, 313–25. Cambridge, MA: Academic Press.

Currie, Adrian. 2018. "From Models-as-Fictions to Models-as-Tools." *Ergo* 4 (27): 759–81.

Elliott, Kevin, and Daniel McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81 (1): 1–21.

Frigg, Roman, and James Nguyen. 2017. "Models and Representation." In *Springer Handbook of Model-Based Science*, ed. Lorenzo Magnani and Tommaso Bertolotti, 49–102. New York: Springer.

Giere, Ronald N. 1988. *Explaining Science: A Cognitive Approach*. Chicago: University of Chicago Press.

———. 2004. "How Models Are Used to Represent Reality." *Philosophy of Science* 71 (5): 742–52.

———. 2010. "An Agent-Based Conception of Models and Scientific Representation." *Synthese* 172:269–81.

Goel, Vinod, and Peter Pirolli. 1992. "The Structure of Design Problem Spaces." *Cognitive Science* 16:395–429.

Haasnoot, Marjolin, W. P. A. van Deursen, J. H. A. Guillaume, J. H. Kwakkel, E. van Beek, and H. Middelkoop. 2014. "Fit for Purpose? Building and Evaluating a Fast, Integrated Model for Exploring Water Policy Pathways." *Environmental Modelling and Software* 60:99–120.

Hilpinen, Risto. 2011. "Artifact." In *Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University. https://plato.stanford.edu/archives/win2011/entries/artifact/.

IPCC (Intergovernmental Panel on Climate Change). 2013. *Climate Change 2013: The Physical Science Basis*. Cambridge: Cambridge University Press.

Jacquart, Melissa. 2016. "Similarity, Adequacy, and Purpose: Understanding the Success of Scientific Models." PhD diss, University of Western Ontario. https://ir.lib.uwo.ca/etd/4129.

Jakeman, A. J., R. A. Letcher, and J. P. Norton. 2006. "Ten Iterative Steps in Development and Evaluation of Environmental Models." *Environmental Modelling and Software* 21:602–11.

Katzav, Joel. 2014. "The Epistemology of Climate Models and Some of Its Implications for Climate Science and the Philosophy of Science." *Studies in History and Philosophy of Modern Physics* 46:228–38.

Knuuttila, Tarja. 2011. "Modeling and Representing: An Artifactual Approach." *Studies in History and Philosophy of Science* A 42 (2): 262–71.

Lenhard, Johannes, and Eric Winsberg. 2010. "Holism, Entrenchment, and the Future of Climate Model Pluralism." *Studies in History and Philosophy of Science* A 41 (3): 253–62.

Levins, Richard. 1966. "The Strategy of Model-Building in Population Biology." *American Scientist* 54 (4): 421–31.

Lloyd, Elisabeth A. 2010. "Confirmation and Robustness of Climate Models." *Philosophy of Science* 77 (4): 971–84.

Mäki, Uskali. 2009. "MISSing the World: Models as Isolations and Credible Surrogate Systems." *Erkenntnis* 70 (1): 29–43.
Mayo, Deborah G. 1996. *Error and the Growth of Scientific Knowledge*. Chicago: Chicago University Press.
———. 2018. *Statistical Inference as Severe Testing*. Cambridge: Cambridge University Press.
Morrison, Margaret, and Mary S. Morgan. 1999. "Models as Mediating Instruments." In *Models as Mediators*, ed. Mary S. Morgan and Margaret Morrison, 10–37. Cambridge: Cambridge University Press.
NRC (National Research Council). 2007. *Models in Environmental Regulatory Decision Making*. Washington, DC: National Academies.
Oberkampf, William L., and Christopher J. Roy. 2010. *Verification and Validation in Scientific Computing*. Cambridge: Cambridge University Press.
Oreskes, Naomi, Kristin Shrader-Frechette, and Kenneth Belitz. 1994. "Verification, Validation and Confirmation of Numerical Models in the Earth Sciences." *Science* 263 (5147): 641–46.
Parker, Wendy S. 2009. "Confirmation and Adequacy-for-Purpose in Climate Modeling." *Aristotelian Society Supplementary Volume* 83:233–49.
———. 2010. "Scientific Models and Adequacy-for-Purpose." *Modern Schoolman: A Quarterly Journal of Philosophy* 87 (3–4): 285–93.
———. 2011. "When Climate Models Agree: The Significance of Robust Model Predictions." *Philosophy of Science* 78 (4): 579–600.
———. 2015. "Getting (Even More) Serious about Similarity." *Biology and Philosophy* 30 (2): 267–76.
Rice, Colin. 2019. "Models Don't Decompose That Way: A Holistic View of Idealized Models." *British Journal for the Philosophy of Science* 70 (1): 179–208.
Rykiel, Edward J. 1996. "Testing Ecological Models: The Meaning of Validation." *Ecological Modeling* 90:229–44.
Schmidt, Gavin A., and Steven Sherwood. 2015. "A Practical Philosophy of Complex Climate Modelling." *European Journal for the Philosophy of Science* 5 (2): 149–69.
Taper, Mark L., David F. Staples, and Bradley B. Shepard. 2008. "Model Structure Adequacy Analysis: Selecting Models on the Basis of Their Ability to Answer Scientific Questions." *Synthese* 163 (3): 357–70.
Teller, Paul. 2001. "Twilight of the Perfect Model Model." *Erkenntnis* 55:393–415.
Thacker, Ben H., S. W. Doebling, F. M. Hemez, M. C. Anderson, J. E. Pepin, and E. A. Rodriguez. 2004. "Concepts of Model Verification and Validation." Technical Report LA-14167-MS, Los Alamos National Laboratory.
van Fraassen, Bas C. 2008. *Scientific Representation*. New York: Oxford University Press.
Weisberg, Michael. 2013. *Simulation and Similarity*. New York: Oxford University Press.
Wimsatt, William C. 2007. *Re-engineering Philosophy for Limited Beings*. Cambridge, MA: Harvard University Press.