Validation: A Window into Economic Practice

A Study into the Practice of Macroeconomic Modeling

Validatie: Een venster op de economische praktijk

Een onderzoek naar de praktijk van het macro-economisch modelleren
(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

vrijdag 29 september 2023 des middags te 12.15 uur

door

Sebastiaan Benjamin Tieleman

geboren op 5 april 1993
te Leeuwarden

Promotoren:
Prof. dr. M.J. Boumans
Prof. dr. J.G.M. van Marrewijk

Beoordelingscommissie:
Prof. dr. K. Frenken
Prof. dr. W.H.J. Hassink
Prof. dr. T.T. Knuuttila
Prof. dr. ir. C.W. Oosterlee
Prof. dr. B.M.J. Rijkers

Acknowledgments

First and foremost, I would like to express my gratitude to my PhD Supervisor, Marcel Boumans, for his support and guidance throughout my research journey. As I completed my Master's degree at Utrecht University, my primary focus was on macroeconomics, and I thoroughly enjoyed delving into models, programming, and data analysis. However, I soon found myself contemplating methodological questions on a deeper level, which led me to approach Marcel about writing a PhD proposal. Our shared interest in macroeconomic methodology solidified my belief that this was the right path to pursue.

Embarking on my PhD, I encountered a significant transition from applied economics to methodology and the philosophy of science, which proved more challenging than I had initially anticipated. I often found myself uncertain about my ideas, but Marcel served as an exceptional mentor during this formative stage, helping me navigate this new field of study. Rather than simply confirming or dictating my thoughts, Marcel encouraged me to develop and explore my own ideas. When I reflect on my years as a PhD student, I recognize them as profoundly transformative, significantly shaping my understanding of economics and science as a whole. For this, I am very grateful to Marcel.

Secondly, I extend my gratitude to my colleagues at the Utrecht School of Economics and the Center for Complex Systems Studies. Being part of such a dynamic and vibrant research community has been an enriching experience, and I have greatly valued the stimulating intellectual exchanges and collaborative atmosphere.

Thirdly, I would like to express my appreciation to my family, whose support has been instrumental throughout my entire educational journey. I am grateful to my brother, who I have looked up to from my earliest years. My mother for her belief in my potential and her ceaseless encouragement to strive for excellence. Lastly, to my father, whose example as a professor serves as a guiding light for my own aspirations. Their collective influence and support have been invaluable in my academic endeavors.

Lastly, but certainly not least, I want to express my deepest gratitude to my fiancée, Natasja, who has been by my side throughout the entire duration of my PhD. I am truly grateful for your unwavering support during the long and sometimes challenging process. Furthermore, I value you as an intellectual partner, as our discussions have enriched my research and provided me with invaluable insights.

Contents

Chapter 1

Introduction

This dissertation aims to provide insights into the methodology of macroeconomic modeling as observed in practice. This first chapter introduces this aim, clarifies why it deserves investigation, and discusses the general approach of the studies in this dissertation.

Section 1.1, offers an explanation as to why models are the main tool of investigation in macroeconomics. Section 1.2 provides an epistemological description of scientific models in the context of macroeconomics. Section 1.3 presents a brief discussion of methodological controversies within macroeconomic modeling practice. Section 1.4 puts forward the various research questions that this dissertation seeks to answer. Finally, Section 1.5 presents the methodology for answering these research questions.

1.1 The Central Problem of Macroeconomic Science

Like most scientific disciplines, the main aim of macroeconomics is to gain an understanding of a particular domain. In most cases, this domain is a system that is thought to operate in our observable world – in other words, a real-world system. The real-world system that is the target of an investigation can be called the target system. For macroeconomics, this is the macroeconomic system, which is constituted by macroeconomic entities. These entities can be divided into two kinds: macroeconomic variables and macroeconomic relationships. Macroeconomic variables are constructed by aggregating the economic characteristics of groups of individual agents. Typical examples of such variables are inflation, interest rates, unemployment, and gross domestic product (GDP). Macroeconomic relationships broadly concern the associations between macroeconomic variables. An example of this is the Philips curve , which describes how higher inflation is associated with unemployment (Phelps, 1967). Such relationships are also of primary interest from a policy perspective. For instance, a policymaker could ask: Will an increase in inflation rate in the economy through lowering interest rates result in a reduction in unemployment? Together, the variables and relationships outlined above make up the system structure.

Having defined the system of interest, the next question is how to study it. Answering questions like the one posed above requires knowledge of the essential variables and relationships within the system. In most scientific disciplines, and especially in natural sciences, the dominant view among practitioners is that true understanding comes in the form of knowledge of the causal relationships within the system. Causal relationships can be formulated as a necessary conjunction between variables (Hoover, 2001), such that if we intervene and change one variable, a necessary change occurs in another. As a result, the fundamental system that is to be understood within such a view is that of a causal structure.

If the structure in question is approached as a causal structure, the preferred method for gaining an understanding thereof is to examine the system within an experimental arrangement that allows for both induced perturbation and control. The most obvious arrangement that fulfills these criteria would be a laboratory setting.

To consider the criteria of perturbation and control in more detail, let us say that we are interested in studying the effect of a change in factor X on factor Y. First, note that to establish whether such an effect exists, there must be variation in the level of X. If X does not vary, there is no way to establish whether X affects Y. In this way, an experimental arrangement allows for the perturbation of X; the research can induce variation in X and then observe what happens to Y.

Second, there may be a third factor, Z, in the causal structure of interest. If we assume that Z has a causal effect on Y, this poses a potential problem. If we change X but at the same time a change in Z also occurs, we may mistakenly attribute the effect on Y to X. This problem is exacerbated if Z also has a causal relationship with X. In this case, each time we induce a change in X, a change in Z will also occur, causing a change in Y. For this reason, in experimental arrangements such as laboratories, emphasis is placed on keeping other factors that may affect Y as constant as possible while studying the effect of X on Y.

Generally speaking, the macroeconomic system cannot be studied within an experimental arrangement such as a laboratory because the scale of the system is large (i.e., economies of entire countries) and the relevant time horizons are long (typically months up to decades). The external validity of laboratory or field experiments in macroeconomics would therefore be too low to provide us with adequate understanding. For most questions relevant to macroeconomics, we cannot recreate the macroeconomic system in an experimental setting in such a way that it still yields understanding about the real economy. We also cannot induce variation in macroeconomic variables in a way that would avoid huge potential costs. Accordingly, inducing variation in variables of interest just for the sake of research is not an option.

It should be noted that valuable research involving experiments does exist in macroeconomics. The most well-known is the work on expectation formation (Hommes, 2011). However, this is most often used to present so-called boundary cases; if agents are unable to form rational expectations in a simple laboratory setup on short time horizons, we cannot expect them to do so in the real economy.

Given this context, the central problem of macroeconomic science is how best to study the structure of a system that cannot be controlled and that can only be passively observed. It is a question that macroeconomists have explicitly struggled with for over a century, and to which they have provided us with several different answers. Since the early 1930s, however, with the work of Michal Kalecki, Jan Tinbergen, and Ragnar Frish, among others, the answer to the question has most often been the use of mathematical models (Boumans, 2005).

1.2 What are Models?

What are mathematical models, and how could they potentially serve to analyze the target system? Multiple accounts of models exist in the philosophy of science. What most of these accounts have in common is that, in order for a model to enhance our grasp of a target system, the model must, in some form and to some degree, be a representation of the real-world structure of interest. The general idea is that rather than studying a real-world structure that we cannot control, we construct representations thereof that we can. By studying the representation, we enhance our understanding of the target system's structure.

1.2.1 Models as Mediators

The account on models that stands as the starting point for the analysis in this dissertation is presented in "Models as Mediators" (Morgan & Morrison, 1999). This account has become a central building-block in my research because I have found it to be a close match to the way in which models are constructed and used in practice.

The "Models as Mediators" account centers around the idea that models integrate three distinct elements: representations of theory, representations of what is empirically observed, and non-representational elements (called artifacts). The structure of a model is thus shaped by the existing theory about how the target system is structured, by the data generated by the target system, and by artifacts.

1.2.2 The Representational Elements of Models

Let us first discuss the representational elements of these models. What should we understand by representation? It is useful to introduce the mathematical concept of mapping as a starting point. A mapping refers to an operation that takes an element in one domain (A) and transforms it to an element in a different domain (B). It is thus about defining relationships between elements in distinct domains. In this sense, a representation can be understood as a mapping between elements in the domain of a specific part of the model structure and elements in the domain of a particular theory or data characteristic.

Mathematically speaking, mapping can be injective or noninjective. Injective here refers to a one-to-one relationship. This means that each element in B corresponds to a transformation of at most one element in A. In non-injective mapping, at least one element in B corresponds to multiple elements in A.

In the context of representation in modeling, injective and noninjective should be understood as extremes presenting a continuum. Some representational relationships may be characterized as being more injective, meaning that a more or less one-to-one relationship is evident between the individual parts of the model structure and the individual elements of theory or data. This implies that the loss of information is limited. The information within the theory or data characteristic overlaps with the information in the model structure to a large degree (mathematically speaking, such a mapping would be invertible). We can, for example, think of the theory of supply and demand, which is represented by a model of supply and demand equations, or the value of the consumer saving propensity parameter in a model quantified by using microeconomic survey data on consumer behavior.

Representational relationships can also be characterized as leaning more toward non-injective. This means that multiple conceivable structural model elements correspond with a particular element of a theoretical concept or data characteristic. In turn, this implies that the theory or data is consistent with but underdetermined by the structural model element. Such non-injective representational relationships are prevalent when the theory or data characteristic only applies to the model structure as a whole. For instance, Calvo pricing is consistent with Keynsian theory, as its function in the model as a whole is to introduce nominal rigidities. However, it is not an injective representation, as many conceivable structures exist that could introduce nominal rigidities. Nominal rigidities can also be understood as something different from Calvo pricing. The same is true for the output data generated by the model. Structural elements of the model that ensure that the model output is consistent with the empirical data can, in most cases, be understood as non-injective representations.

Importantly, as I will return to later, non-injective representations cannot be understood as purely representational mappings. Because the relationship between the model's structural elements and the object to be represented is, by definition, underdetermined, choices must be made that are not primarily informed by that object. Structural elements that are non-injective representations of theory, for example, can also be non-injective representations of data; or, as we will see is often the case, they may be artifactual.

The term "injective" is normally used within a strict mathematical context. It is my view, however, that the term is instructive here because of its analogous meaning. Additionally, etymologically speaking, injective comes from "inject," meaning to force or drive something into something else. This corresponds to the stricter and less flexible way in which theoretical notions or data characteristics are incorporated into model structures in the case of injective representations.

Additionally, note that for both injective and non-injective representations, the relationship generally has a partial nature. Given the complexity of the system of interest, models always involve omissions. This implies that a selection has to be made relative to the specific purpose of a model. In the case of macroeconomics, we may not be interested in studying all the macroeconomic relationships at once; rather, we are specifically interested in the relationship between inflation and unemployment, for example. The model construction process therefore requires a selection of theoretical and empirical elements to be integrated into the model that are relevant for this specific relationship. This omission can come in the form of simplification, which involves the omission of details of a representation or that of the theory or data from the model altogether.

1.2.3 The Artifactual Elements of Models

Let us now discuss the third element from which models are constructed: artifacts. Artifacts are elements of the model that are representative of neither the theory nor the data. They are most often incorporated into the model simply to make it work relative to some intended purpose. The term “work” here could refer to the mathematical properties of a model. We can think, for example, about the infinite time-horizon optimization problems that agents in most modern macroeconomic models are assumed to solve. Neither theory nor data dictate that agents solve an optimization problem that involves an infinite time horizon. Concepts such as infinity (the same is true for zero) often have desirable mathematical properties. They allow the model to be solved analytically, for example.

It may also be that the model’s purpose is to show different possible structures without claiming that they are directly representative of the real-world structure. Knuuttila (2021) provides us with the example of the ultra-Keynesian model constructed by Tobin (Tobin, 1970). The purpose of this model was to counter Friedman’s conclusion that the level of money supply was the only source of inflation. Friedman based this conclusion on a particular empirically observed conjunction of the variation in the time-series data of money supply and inflation. Tobin’s model was able to show that a model in which money played no causal role was also able to reproduce the time-series data pattern. Tobin did not seek to claim that the structure of his model was an injective representation of the real world but rather to show that the time-series data alone were not sufficient evidence to show that money supply was the only driving force behind inflation. Since the output of the model as a whole was consistent with observed time-series data, Tobin’s model structure was a non-injective representation of that dataset. There would have been many possible structures consistent with the observed phenomenon, which was precisely what Tobin aimed to demonstrate. Given that the structure of Tobin’s model cannot be inferred from data alone, it can be characterized as artifactual to a large degree.

We have discussed the three main elements from which models may be constructed: representations of data, representations of theory, and artifactual elements. An important complication is that, in many cases, it may not be so clear whether a particular element is representational or artifactual. Indeed, the elements themselves may often be an integration of both representational and artifactual aspects. As discussed in the case of Tobin's model, this is particularly the case when representations are non-injective.

Alternatively, artifactual elements that turn out to be crucial to make the model work may be justified by practitioners on theoretical grounds. An example is Calvo pricing (Calvo, 1983), which is a crucial element in most new Keynsian dynamic stochastic general equilibrium (DSGE) models to induce nominal rigidities. Calvo pricing entails that a firm has a constant probability of being able to reset its price independent of when it was last able to do so. While it has been shown that this structural element is empirically false (Álvarez & Burriel, 2010), attempts have been made to show that Calvo pricing is consistent with microeconomic theory following its popularization in new Keynsian DSGE models (Woodford, 2009).

This discussion of artifacts fits within a larger discussion in the philosophy-of-science literature on how best to characterize scientific models. Knuuttila (2021) forwards a view that is close to the general view presented here and therefore also closes the account by Morgan and Morrison (1999). It is a view that contrasts with a more traditional view that puts all the epistemic value of the model in the representative relationship between the model structure and the real-world structure. Knuuttila and Loettgers (2016) does not dispute that what is labeled as external representation can be a source of epistemic value for models; rather, given the various distinct purposes a model may have, it is not the only way to characterize the epistemic value of models.

The view of models present throughout this dissertation is open to this broader view of their epistemic value. In fact, as I will argue throughout, it is highly questionable whether the structure of macroeconomic models can be credibly conceived as an injective representation of the real-world structure, nor is it clear that this is the purpose of the model builder in most cases.

The way in which I discuss the nature of artifactual elements here is to a large extent in line with Herbert Simon's discussion in "The Sciences of the Artificial" (Simon, 1969). In this work, something that is characterized as artifactual (or artificial) is constructed by human beings, imitates the appearances of natural things only in limited domains, and is heavily characterized by its purpose. In Hoover (1995), we find an account that relates this idea to models in macroeconomics. Crucial here is the relative independence between the inner environment of the model and the real-world outer environment in which it operates. An artifactual model is correct if its inner environment is appropriate from its outer environment. An insightful example here is that of a clock (which can be read here as analogous to an economic model), the function of which is to tell the time. While two clocks may indicate the time in the same way, it is conceivable that they may be constructed differently. In other words, whereas their inner structure is different, for the purpose at hand (to tell the time in the outer environment), they are both functional. A degree of independence thus exists between the inner and outer environments. Note that this is in line with our description of non-injective representations. We will return to this discussion in Chapter 6.

1.2.4 Integration through Mathematical Molding

Let us now look more closely at the integration process itself. How are data, theory, and artifactual elements integrated into a uniform model object? Let us first establish that models are constructed from materials of various types. Human anatomy models, for instance, are often made from plastic and paint. Models can be graphical, in which case they are constructed from lines and curves. In the case of macroeconomic models, this material is often a particular type of mathematics. A DSGE model, for instance, is a system of difference equations. The material is what allows the integration of the various model elements in a process that has been labeled as mathematical molding (Boumans, 1999). Importantly, the type of material and the process of mathematical molding present constraints for the model builder within which the model is to be constructed. In particular, the incorporation of some artifactual elements can be indirectly necessitated by the type of material chosen.

1.2.5 Models as Autonomous Objects

Now that we have established the elements of which models generally consist, we can ask why they are useful. Within this account, the utility of models comes from the fact that they are mediators between theory, data, and artifactual elements. That is, through a process of mathematical molding, models are able to integrate representations derived both from theory and from data together with artifactual elements into a single object. This object can then be studied and experimented on.

It is essential to note here that models are not merely an extension of theory and/or data, which are traditionally considered the primary devices of knowledge production.

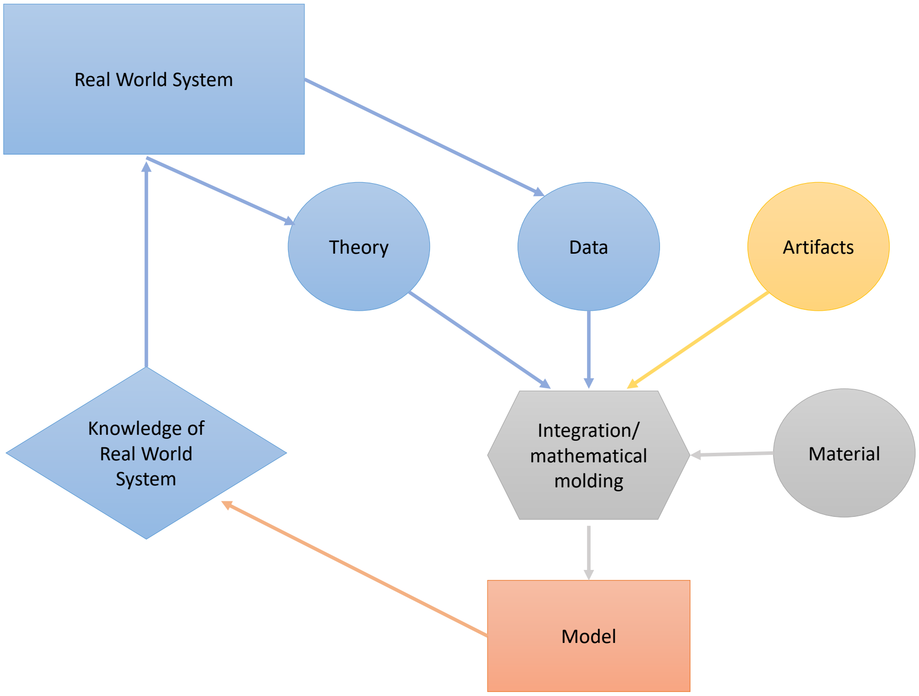

Real World System
Theory
Data
Artifacts
Knowledge of Real World System
Integration/ mathematical molding
Material
Model

Figure 1.1: Models as Mediators

The process of model construction that we have described above is one of integration through mediation, the outcome of which is a model that may be dependent on theory and/or data, but also to some degree independent from them. Modeling requires making particular structural choices that are not given by theory and data. As concluded in Morgan and Morrison (1999), models are therefore autonomous objects.

The fact that models are autonomous objects implies that they are a worthy subject within philosophy-of-science research. Hence, an investigation into how models are constructed, validated, and applied has the potential to answer methodological and epistemological questions. This potential is especially significant for scientific domains that rely on the use of models to a large degree, such as macroeconomics.

In Figure 1.1, we can see a schematic summary of the model account discussed. To begin, models are constructed from three types of elements: representations of the theory we have about how the real-world system operates, representations of data generated by the real-world system, and artifacts that have no relationship with the real-world system. These elements are integrated into a process of mathematical molding. Integration requires the choice of a particular building material, such as a type of mathematics. The outcome of this process is a model. By studying the model's structure and by observing how it functions, we acquire an understanding of it, which can then be translated into knowledge of the real-world system. As a result, the model serves as an instrument to acquire knowledge of the real-world system. It does so by mediating between various pieces of information on which we can draw about the real-world system.

1.3 A Lack of Consensus in Macroeconomics?

Thus far, we have discussed how, due to the impossibility of studying the macroeconomic system in an experimental environment, the primary tools used in macroeconomics are mathematical models. We have discussed in general terms how models are an integration of different types of representational elements and artifacts.

However, this general description of a model does not in itself tell us which combination of elements produces a model that is the right tool given a particular question about the macroeconomic system. Historically, various schools of thought have been forwarded regarding which type of macroeconomic model is to be preferred (Hendry, 2020). Different schools of thought have had different preferences regarding which theoretical and empirical representational elements should be incorporated into the model and which mathematical form is desirable.

Duarte (2012) portrays a picture of a macroeconomic science that is characterized by periods of both controversy and consensus. In recent decades, a consensus has been formed around various versions of DSGE models. However, many practitioners also perceive macroeconomics to contain a sharp division between mainstream and heterodox approaches (Lee, 2012). While the correctness and usefulness of this assertion warrants an investigation on its own, it is true that distinct methodological approaches exist when it comes to macroeconomics modeling, and that practitioners specializing in one modeling approach often criticize the methods of others.

An example of such distinction between approaches are agent-based modeling practitioners, who often formulate fundamental methodological criticism of DSGE models. On the other hand, economists specializing in DSGE approaches are reluctant to change their methods in response to such criticisms (Hoover, 2021). While most criticisms of DSGE models have been around for some time, most of them gained traction after the 2008 financial crisis (see Stiglitz (2011) for an overview of several criticisms). One source of criticism arises from a heterodox approach known as agent-based macroeconomics Farmer and Foley (2009). Agent-based macroeconomics emphasizes the role of interactions between many distinct agents in generating certain macroeconomic patterns. It posits that because DSGE models did not take such interactions into account, they were unable to reproduce prolonged periods of economic decline, such as those observed after the 2008 crisis. I will elaborate on this criticism in Chapter 2.

What is important for the discussion here is that macroeconomic agent-based models (MABMs) are constructed from different representational elements than DSGE models. This is true in terms of both theoretical elements (complexity theory and behavioral economic theory versus general equilibrium theory) and empirical elements (the reproduction of prolonged downturns versus return to equilibrium). Since the representational elements are different in both approaches, the process of mathematical molding is also different. Basic DSGE models can be solved analytically and rewritten into three difference equations (Woodford, 2003). Analytically, solvability is generally not a requirement for MABMs, which are instead studied by using Monte Carlo simulations.

DSGE models and MABMs are thus constructed from contrasting methodological perspectives. Interestingly, however, both approaches are used to answer similar questions. Methodological disagreements of this kind are common in macroeconomics; as such, they point to the fact that most methodological discussions within macroeconomics are by no means settled.

We have established that models occupy an autonomous space within science and that fundamental disagreements are apparent among practitioners over how best to construct models in macroeconomics. The study of macroeconomic models is, therefore, not only interesting intrinsically from a philosophical perspective but also highly relevant for practitioners of macroeconomics. Methodological reflections, including those presented in this dissertation, can help practitioners to better understand the foundations of their own tools and potentially guide the direction of future research.

1.4 Aims and Research Questions

The subject of this dissertation is the model construction process in macroeconomics as observed in scientific practice. The central aim is to provide a more systematic understanding of the model construction process. As will be discussed in greater depth in the following section, the methodology of this dissertation is primarily inductive. This implies that the starting point of this investigation is the observed scientific practice in the form of various cases of model construction and their use. In turn, the insights derived from these cases will provide input into a more general framework of model construction.

Three distinct cases will be the subject of this investigation. These three cases also represent the main chapters of this dissertation: the introduction of macroeconomics agent-based models (Chapter 2), the hybrid model critique of DSGE models (Chapter 3), and inter-domain model transfers (Chapter 4).

Macroeconomic Agent-based Models[1]

The use of MABMs is a relatively new approach in macroeconomics. Their use gained popularity and recognition after the 2008 financial crisis. However, open questions remain among macroeconomists about the epistemic role of such models and how they are validated.

The Hybrid Model Critique on Dynamic Stochastic General Equilibrium models

DSGE models generally consist of a structural core and a stochastic periphery. This has received criticism because it is unclear how such a model structure relates to empirical data. A fundamental analysis of the criticism is still lacking in the current literature.

Inter-Domain Model Transfer[2]

Inter-domain model transfer refers to the transfer of models between different scientific domains. Models originally developed in biology, for example, are in some cases re-used in economics. However, open questions and unexplained observations remain regarding particular cases of model transfer.

[1] This chapter is a reworking of the paper that I published in the journal Computational Economics (Tieleman, 2021). The core of the analysis in this chapter and the paper are the same. The order of sections, formulations and other details may have been altered.

[2] This chapter is a reworking of the paper that I published in the journal Synthese (?, ?). The core of the analysis in this chapter and the paper are the same. The order of sections, formulations and other details may have been altered.

Model Construction Framework and Concluding Remarks

The various insights from these cases can be distilled into a framework of model construction, which will discussed in Chapter 5. The concluding remarks in Chapter 6 will provide a summary of the dissertation along with a general discussion of modeling practices in macroeconomics.

1.5 Methodology

How should we go about answering the research questions formulated in the previous section? I will now clarify the methodological approach taken in this dissertation and argue why it is well suited to the purpose at hand.

1.5.1 Philosophy of Economic Science-in-Practice

Generally speaking, to gain a more fundamental understanding of science means to step into the domain of the philosophy of science. However, multiple routes can be taken therein in the pursuit of understanding. Boumans and Leonelli (2013) distinguish between two such routes. The first is labeled as philosophy-of-science in practice – a more traditional route in which one starts from certain formal philosophical accounts, such as that of scientific explanation, before seeking to fit a scientific method as observed in practice to this formal account. In this way, the formal account serves as a benchmark for assessing whether a scientific method yields a correct form of knowledge production. As it turns out, a perfect match is never achieved between the practice as it is observed and how science should be according to philosophers of science.

The value of this type of investigation lies in the fact that it is able to drag a scientific practice outside of its internal confines and judge it through external criteria – that is, in the sense that they are to some extent formulated independently from what is observed within scientific practice. Whether these external formal criteria actually provide a deeper understanding of the scientific method is dependent on one's epistemological and ontological views.

The limitation of this approach, as is also argued by Woody (2014), is that formal accounts are not just a benchmark but also serve as a filter. Establishing how a particular scientific practice functions relies on empirically observing it. With any empirical observation also comes selection; some information is considered relevant, while other information is left out. The formal philosophical accounts guide this selection process. When looking at scientific practice through the lens of formal explanation, for example, little room is generally available for an investigation of the historical context within which scientists came to their conclusions. The same is true for an investigation into the scientific paradigms within which scientists operate and their interactions with peers. Such information is generally irrelevant from the perspective of the formal philosophy of science accounts. A potential problem with this is that one misses information that is crucial to providing an empirically accurate account of scientific practice.

Woody (2014) provides a case study of the discovery of the periodic table. It shows that knowledge about the interactions between scientists in particular is crucial to understanding how the period table was actually developed. Analysis of the periodic table based on formal philosophical accounts has generally not taken this into account.

The second route, as discussed by Boumans and Leonelli (2013), is labeled the philosophy of science-in-practice. Here, one does not start with a particular formal philosophical framework in mind; instead, one tries, as neutrally as possible, to start from the data. That is, to observe scientific practice along a broad range of facets, including those elements generally absent from formal philosophical accounts. The focus is on understanding the internal structure of a scientific practice without necessarily seeking to judge the practice from the perspective of a formal philosophical account. The advantage of this approach is that it stays close to scientific practice in terms of how it is actually observed. In this way, it is open to facets that may be highly relevant to understanding how scientists work.

A criticism of this approach is that it may appear circular to some – understanding science by understanding science. However, such criticism presumes that understanding science-in-practice is a trivial matter. This is not the case. The choices that practitioners make, and why certain methods gain traction while others do not, are often not discussed explicitly and publicly by the practitioners themselves. In some cases, the practitioners are unaware of why they make certain choices or why they appreciate certain research studies more than others. They take part in scientific paradigms that bring with them certain intuitions and implicit judgments. The added value of the philosophy of science-in-practice is to open this black box of intuitions and make the implicit more explicit. In turn, this allows scientists to reflect on their own methodologies and guide future research.

The approach adopted in this dissertation is that of the philosophy of science-in-practice. My goal is to understand how macroeconomic models are constructed in practice and to open up the black box that is the model construction process.

The next question is how we should go about doing this. Philosophy of science-in-practice is an inductive form of research. That is, it seeks to make more general statements by observing and analyzing multiple individual cases. However, analyzing scientific practice based on observation is not trivial.

To begin, a relevant question is how to observe science in practice. Given that the goal of philosophy of science-in-practice is to make methodological considerations explicit that are often implicit or intuitive, I have found three main sources for making observations as useful as possible. The first is to conduct interviews with practitioners, asking them why they made certain choices; much can be often learned from such activity. The second is to trace back a certain method to its original formulation. The third is to study the methodological reflections of practitioners themselves if they are available.

Any form of scientific observation, however, must rely on a useful interpretive framework that gives meaning to the observations. In the next subsection, I argue that model validation provides a basis for such an interpretive framework.

1.5.2 Model Validation as an Interpretive Framework of Science in Practice

To give our observations meaning, one cannot escape the use of some interpretive structure, which can come in the form of a framework. A framework filters and transforms observations into a coupling of observations and interpretation. It is this coupling that is useful and, when brought together with couplings from other cases, it can be used to induce more general statements about science-in-practice.

The question then becomes: What is a useful framework for analyzing science in practice and, in particular, the construction and use of economic models? First, it must be stressed that a pluralist attitude to such a question is, in my view, most useful. Each framework highlights certain elements and discounts others, and no framework is all-encompassing. Many of the frameworks that have been developed in the philosophy-of-science of economic models provide valuable and unique insights;

in this research, I often build on such contributions. To acquire balanced insights into scientific practices, studying them through several interpretive frameworks is essential. If we envision economic scientific practice as a complex system that happens inside a closed-off house, each useful interpretive framework is a window into its interior workings.

In the chapters of this dissertation, I will demonstrate that using model validation as the center point of such an interpretive framework yields useful and novel insights into various modeling practices. In other words, it represents a large window with clear views into the house of economic scientific practice.

Before demonstrating this, however, let us first discuss why model validation is such a useful concept for studying scientific practice. Model validation is the assessment of a model's ability to fulfill its intended purpose. It is a crucial element in the model construction process that scientific practitioners address both implicitly and explicitly. At this point, several elements require unpacking to clarify why and how I make use of the concept of validation to understand scientific practice in the context of modeling.

First, the definition presumes that models are constructed to fulfill some intended epistemic purpose. That is to say, models are constructed for specific reasons. In some instances, such reasons are explicitly stated. In other instances, the reasons why models are constructed are to be inferred from a particular research context or paradigm. Often, models are intended as tools to answer a particular research question. I have found that different questions can be related to particular forms of validation. This implies that by studying how models are validated in a broad sense, the purpose of a model often becomes clear. Is the purpose to provide a scientific explanation of a phenomenon, or is it the precise measurement of some fact or other? In addition, we can think of other types of purposes, such as positioning one's model within a particular scientific paradigm. Such purposes may also become clear through studying how a model is validated.

Second, model validation not only requires revealing the model's purpose but also sheds light on how such a purpose is best fulfilled by the model from a more methodological perspective. Model validation involves the formulation of what I label validation criteria. Validation criteria are points of reference by which the performance of a model can be assessed relative to its purpose. These points of reference are operationalized to be of direct use in the model construction context; the criteria entail what the model should and should not be able to do. This step thus reformulates the epistemic model purpose as methodological criteria. By investigating which validation criteria apply, the relationship between the epistemic purpose of the model and the method of fulfillment can be better understood.

Validation is thus a useful element in an interpretive framework because it brings both the epistemological and methodological convictions of the model builder into the foreground. In this sense, it has the potential to provide an integral understanding of the model construction process from the more fundamental aspects to the specific technical choices made. In itself, however, this does not ensure that validation as a basis for an interpretive framework enables us to make observations that lead to this type of understanding. In most cases, only a few of the epistemological and methodological convictions will be explicitly stated by the model builder. I will argue that validation has an additional characteristic that helps to mitigate this obstacle; validation involves a type of judgment. That is, the validation process assesses how the model performs in the fulfillment of a wide range of criteria. This judgment is immanent, meaning that the criteria by which the model is assessed are internalized within the model construction process because they are given by the model purpose.

To explain this further, let us recall that model validation is defined as the assessment of the model's ability to fulfill its intended purpose to a sufficient degree. The result of a finalized validation process is thus a judgment: the model either fulfills its purpose, in which case it is successful and the results can be published, or it does not and the model is to be discarded or adjusted. This judgment extends to the elements of the model itself, implying that the combination of model elements is seen as optimal for the purpose at hand given the constraints to which the model builder is subjected. A different way to describe this conditional optimality is that the model is epistemically or methodologically preferred to other hypothetical model configurations given the same constraints. This presents us with what can be labeled as a normative ranking: given the same constraints and purpose, Model A is preferred to Model B, for example. Based on such observations, epistemological and methodological convictions can be indirectly inferred.

This ranking provides us with useful and clearly defined pieces of information. For example, in Chapter 6, I will argue that agent-based models are constructed from a more realist perspective on modeling compared to DSGE models, which rely more on an instrumentalist view. I reach this conclusion by studying how both types of models are validated in different ways.

1.5.3 About the Structure of this Dissertation

In the chapters that follow, I will analyze various cases in macroeconomic modeling practice. The starting point of the analysis will be to observe which considerations are at play in the validation process of various macroeconomics models. To investigate the cases through an interpretive framework of validation, it is often necessary to introduce concepts associated with model validation. The cases differ in their emphasis on which associated concepts are useful. For example, the discussion of agent-based models in Chapter 2 places particular emphasis on empirical validation – that is, assessing whether the output of the model is in line with the empirical data.

Accordingly, each chapter that treats a particular case will start with a discussion of the concepts necessary to understand the rest of the analysis in that chapter. Some concepts introduced in the chapter will be particular to the individual case study. Concepts are also introduced in one chapter that will be built upon in a subsequent chapter on a different case. Empirical validation, as introduced in Chapter 2, for instance, will play a role in the analysis of the cases in Chapters 3 and 4. Wherever this is the case, I will explicitly refer to the place within this dissertation where the concept was first introduced.

The fact that concepts are introduced before the analysis of the cases serves to provide logical structure and is not necessarily a reflection of the order of the scientific process. For all the concepts introduced, it holds true that they are arrived at through an integration of conceptual notions present in the existing literature and my own observations of scientific practice. The observation of scientific practice has been mainly through the analysis of literature in which models are constructed and discussed, as well as seminars and a series of interviews with practitioners of macroeconomics. In many cases, the concepts introduced in this dissertation are thus the result of an inductive process of analysis of the cases themselves rather than a predetermined means of analyzing them.

As mentioned above, the concepts introduced in each case will be integrated into a model construction framework in Chapter 5. It is this framework work that can be thought of as the main generalized contribution of this dissertation. It provides a novel view of the model construction process, with the notion of validation at its center.

Chapter 2

Towards a Validation Methodology of Macroeconomic Agent-Based Models

2.1 Introduction

Macroeconomic agent-based models (MABMs) are a promising new tool in the analysis of macroeconomic phenomena (Farmer & Foley, 2009). These models do not rely on ex ante equilibrium assumptions, which makes them particularly suitable for the analysis of economic crises. However, while MABMs have presented several methodological innovations compared to DSGE models, these same innovations have also been the source of criticism over their use. Most of this criticism has focused on how MABMs are empirically validated (Fagiolo, Moneta, & Windrum, 2007). Such criticism is partially due to the relative novelty of MABMs. A deeper understanding of the relationship between empirical validation practices and the structure of MABMs is therefore required to gauge the correctness of these practices .

In this chapter, a methodology for the empirical validation of macroeconomic agent-based models will be presented. Empirical validation here refers to the assessment of a model's ability to answer a question by comparing relevant characteristics of the model with empirical data. The reproduction of relevant empirical characteristics is what we refer to in this dissertation as the fulfillment of phenomenological validation criteria. In this context, the term "phenomenological" underlines that the validation criteria refer to the phenomenon under study.

The main question that will be answered in this chapter is: How do phenomenological validation criteria in MABM work to enhance model validity, in the way that it is observed in practice? In addition, the analyses of MABMs in terms of their validation allows us to shed light on some more fundamental issues regarding our characterization of MABMs.

This chapter is structured as follows. First, I will introduce MABMs as complex systems with emergent properties. Second, I will connect the notion of MABMs as models of complex systems to how MABMs are validated in practice. By way of illustration, the model and the validation approach in Lengnick (2013) will be used as a case study. Third, I will explore the concept of model validation in greater depth by introducing a framework of model validation. Central to this framework will be classifying the model into certain types based on how they are validated. Fourth, I will apply the framework of model validation to the case in Lengnick (2013).

The analysis introduces a validation methodology for MABMs that reveals fundamental insights into the models. It allows us to pinpoint the constituent structure of MABMs. The structural elements at a lower level are distinct from, but inputs to, the higher-level structural elements. Since structural elements at different levels are validated in different ways, I come to a specific characterization of MABMs within the classification of Barlas (1996) and Boumans (2009), that is, in some ways distinct from other types of macroeconomic models.

In addition, in Section 6.4 of the conclusion of this dissertation, I will provide a more in-depth discussion of realist and instrumentalist views on modeling when applied to MABMs.

2.1.1 Discussion of Relevant Literature

This chapter builds on and contributes to several strands of the existing literature. First, some of the literature seeks to explicate the general issues that modelers encounter in the validation of MABMs, as well as discussing the upsides and downsides of the different validation methods used in practice. The first publications in this series were Fagiolo et al. (2005), Fagiolo et al. (2007), and Windrum et al. (2007). Later important updates followed to discuss new developments, namely Fagiolo and Roventini (2017), Gatti et al. (2018). In this series of papers, the most commonly used validation approaches are forwarded.

In Fagiolo et al. (2005), the main validation approaches discussed are qualitative simulation modeling, replication of stylized facts, empirical calibration, and the history-friendly approach. In qualitative simulation modeling, the relationship between the output of the model and the empirical data is only required in a qualitative dimension. That is, as long as the model output is roughly in line with some qualitative, empirically observed features, the model is considered valid. Such models are most applicable for exploratory and experimental purposes.

Replication of stylized facts is the approach in which the model is considered valid if it is able to reproduce a set of relevant (given the model's purpose) stylized facts. Importantly, all the model parameters are calibrated indirectly, meaning that they are quantified such that the model is able to reproduce the set of stylized facts. This approach is labeled as indirect as the parameter values are not taken directly from data; rather, they are selected so that the model as a whole fits certain stylized facts.

The *empirical calibration* approach is similar to the replication of stylized facts approach in the sense that stylized facts are used to calibrate the model parameters. In addition, however, some of the parameters are calibrated directly. As a result, the empirical data used for the calibration concerns the individual relationship in which the parameter occurs, instead of comparing the output of the model as a whole. In the case of agent-based models, most of the parametrization occurs at the micro level, implying that, in direct calibration, empirical data at the micro level are used. This type of validation is considered a stricter type of validation.

Finally, in the *history-friendly* approach, the validation criterion is to reproduce a precise data history, at least in qualitative terms. Often, this comes in the form of reproducing specific time-series data. In this approach, additional importance is given to matching the initial conditions of the model to those observed in the time series to be reproduced.

Most recently, Fagiolo et al. (2019) outline the latest developments regarding validation techniques of MABMs. This includes the use of machine learning techniques to select regions of the model parameter space that exhibit interesting behavior. Such computational techniques have become increasingly important as agent-based models have become more elaborate, and the criterion of sensitivity analysis has become increasingly important. Sensitivity analysis is an assessment of how robust certain model behavior is against changes in the parameter space, or to changes in assumptions.

An additional interesting development discussed in Fagiolo et al. (2019) is the validation approach described by Guerini and Moneta (2017). In this approach, a VAR-estimation is performed on the aggregated (macro) variables of the simulated data (generated by the agent-based model). The coefficients of this simulated VAR-estimation are then compared to those of the same VAR-estimation performed on empirical data. If the coefficients match in terms of their sign and, to some extent, their size, the model passes the validation test. This is an interesting approach because it addresses what is known as the conditional object critique originally put forward by Brock (1999). The core of this critique is that stylized facts may be too general, to the point that multiple distinct models may be able to reproduce them. The approach in Guerini and Moneta (2017) provides a more selective validation measure within this context.

The publications discussed above have been critical in providing overviews of validation approaches in MABMs in order to build a more standardized approach and are thus important in clarifying the directions in which new research should develop. However, this strand of literature has not sought to provide in-depth methodological foundations for the methods they present. Although it provides an overview of some methodological and epistemological issues in model validation, such as realism versus instrumentalism and underdetermination (see, for example, Windrum et al. (2007)), it does not go one step further to look at how these issues specifically apply to MABMs.

An alternative angle by which we can consider the validation of MABMs is to construct a benchmark model, the performance of which serves as a minimum criterion for model validation. In Caiani et al. (2016) and Lengnick (2013), for example, the aim is to present a benchmark or baseline model, but the difficulty with such a common benchmark may be that the actual validation criteria may differ given the various different questions that the models are built to answer. In this way, benchmark models seek to incorporate only features that are seen as essential for most models. Looking at Lengnick (2013), for example, we see that the incorporated features include households, firms, and banks that behave and interact according to simple rules.

The literature discussed above highlights the variety of existing approaches to the model validation of MABMs. Some approaches may be more suitable for particular types of models than others. For the purpose of this chapter, I will focus on the stylized fact (or indirect calibration) approach, because this is the approach most frequently used in practice (Fagiolo et al., 2019), and because more recent approaches (such as Guerini and Moneta (2017)) are ultimately extensions of this approach.

As we will see, the notion of complex systems and complexity economics plays an important role in the analysis of this study. My own understanding of the economy as a complex system has been strongly influenced by works such as Arthur (2013). Furthermore, the concepts of reductionism and emergence in the context of macroeconomic modeling will be relevant. Hoover (2015) and, to some extent, Gatti et al. (2011) are important contributions in this regard. This study will contribute to the extant literature by connecting the concepts of complexity and emergence with model validation.

Finally, this contribution somewhat stands within a strand of the literature that considers the methodological aspects of agent-based modeling in a general sense. The most well-known studies in this regard are Epstein (1999) and Epstein (2006), in which the idea of agent-based models as tools to generate explanation is brought forward. This is also linked to validation, since a necessary condition within this concept is that agent-based models are valid only if they are able to generate an explanation starting from interacting agents. Furthermore, contributions such as Elsenbroich (2012) and Grüne-Yanoff (2009) have helped me to gain a deeper understanding of agent-based methodology in relation to explanation. However, an analysis of validation in light of the fundamental methodology of agent-based models is not present in the current literature, which is where I hope to contribute.

2.2 Complex Systems with Emergent Properties

The first element necessary to arrive at a validation methodology of MABMs is the notion that the structure of MABMs presents a complex system with emergent properties. This notion entails that, generally speaking, MABMs are tools to model the economy as a complex system. But how should we understand this?

A plethora of definitions exist for what constitutes a complex system. One that is useful for our analysis comes from Ladyman et al. (2013): a system in which elements react to the patterns that they together create through interactions. Hence, according to this definition, a complex system is one comprised of a multitude of interacting elements. Through these interactions, patterns are created to which the elements react. This implies that the created patterns are not always consistent with individual behaviors; in turn, it follows that the natural state of the system is not, at least ex ante, equilibrium (Arthur, 2013). This discrepancy between patterns at different levels is known as emergence.

Emergence is generally a key feature of complex systems, but it can be defined more exactly in numerous ways (O'Connor & Wong, 2015). To analyze the role of emergent properties in models of complex systems more fully, I start from the framework introduced in Baas and Emmeche (1997). In their framework, a system is defined in terms of entities E^n, interactions between these entities Int^n, and an observation mechanism Obs^n. Together, E^n and Int^n make up the structure of the system. Obs^n measures the properties P^n of the entities E^n. Obs^n can be seen as observation mechanisms used by an observer external to the system, while n represents the level of interactions under consideration. Entities at a certain level are constituted by the lower-level entities and their interactions. Formally, levels are connected in the following way:

$$E^{n+1} = M(E^n, Int^n, Obs^n), \tag{2.1}$$

where M can be considered as some process of mathematical induction. Each of these levels of entities can have properties; emergence can be defined through the subsequent definition of these properties. A property P is emergent if $P \in Obs^{n+1}$ while $P \notin Obs^n$. This means that a property of a certain entity is emergent if it is not observed as a property of a lower-level entity.

Let us now see how this framework is reflected in models of complex systems, such as MABMs. Such a model always starts from a base level: $n = 1$. This level consists of the elementary entities of the model and their interactions. All model assumptions are at this level. We can generate the next level in such a model by taking into account the consequences of interactions between these entities E^1 through a process of mathematical induction.

Considering the interactions between entities in this model implies that we use some methods of mathematical induction, such as iteration, to generate the entities at the next level. These entities will, in turn, interact to generate the next-level entities, and so on. If an agent at a level higher than $n = 1$ has properties that can only be observed by considering the implications of interactions of lower-level entities, we can say that an entity has emergent properties. For micro-founded modeling macroeconomics, in which the micro level can be seen as $n = 1$, this implies that – contrary to the "strong reductionist" (Gatti et al., 2011) approach of representative agents – macro properties do not reduce to properties of individual households or firms in isolation. Rather, we require the interactions between agents from which macro properties emerge, which are qualitatively different from the properties of the agents.

Most MABMs can best be described as three-level systems: the micro level ($n = 1$), the meso level ($n = 2$), and the macro level ($n = 3$). Interactions at the micro level give rise to the second-order entities, which consist of networks of agents. These networks, in turn, can interact, resulting in the macro level (i.e., third-level entities). Levels are defined based on whether interactions exist between entities at the same level. However, it is important to consider that in reality, the boundaries between the meso and macro levels are much more blurred than the framework of Baas and Emmeche (1997) would consider them to be. Rather, interactions and feedbacks are evident between the different levels, which can make a level in between the agent and the aggregate model output difficult to set apart.

Let us now discuss how this framework relates to the notion of model structure. Broadly speaking, the structure of a model is defined as the model mechanisms that in some way generate the model target. They are the sum of all E^n and Int^n at all levels n. The model target is the phenomenon that the model is constructed to reproduce. If the purpose is to provide an explanation of business cycle dynamics, for example, the model target is the business cycle, and the model structure comprises the mechanisms by which the model is able to reproduce the business cycle.

In the case of MABMs, the model structure arises from the properties of the entities and their interactions. Given that MABMs are structured as multiple levels, their structural elements operate at multiple levels as well. To understand this, let us look again at the framework of Baas and Emmeche (1997) and further clarify the role of the observation mechanism Obs. Obs states at which level we are observing E and Int. The difference between a lower and higher level of Obs is given by whether we take the structural consequences of the interactions into account. But what does this mean, exactly? Let us consider a two-level example, starting from the micro level $n = 1$:

$$E^2 = M(E^1, Int^1, Obs^1) \tag{2.2}$$

Now, if we put on the Obs^1 glasses, we observe the behavioral rules of agents as properties of E^1. Embedded in these rules is the structure that generates the output at the level of the individual agent. Importantly, however, these rules can, and will in practice, also include rules that are a function of variables of other agents. It could be, for example, that a firm sets its price as a mark-down on the average price of a subset of other firms. These types of behavioral rules can still be observed through Obs^1 and can be understood as Int^1. Together, they embed the structure at the first level. Observing the interaction rules by themselves is not the same as taking into account the structural consequences of those interactions. This requires close consideration of the implications of the behavioral rules of agents for each other.

In the price-setting example, accounting for the structural consequences of interactions would require consideration of the price-setting behavioral rules of other firms as well. At the level of the agent in isolation observed through Obs^1, the influences of other agents are exogenous. However, if we take into account the consequences of interactions, we observe through Obs^2 a system of connected agent entities; the exogenous variables at the micro level are now endogenous in relation to the second-order entities. This system, as observed through Obs^2, will thus be constituted as a structure that accounts for the effects of agents' behavior on each other.

We must note at this point that the difference between observed levels is thus merely one of steps in a mathematical induction. Indeed, this is also implied by the fact that the E^2 is generated through an inductive process M. The reason for this is that the interactions Int^1 do not constitute any new inputs into the model; rather, they are an inductive consequence of the behavioral rules that are part of E^1.

Finally, it is important to see how the notion of emergence fits into this distinction between levels. We have defined emergence as $P \in Obs^{n+1}$ while $P \notin Obs^n$; this implies that the properties of E^2 cannot be observed without taking into account the implication of each agent's behavior on one another. In the case of the model structure, this means that by accounting for the consequences of interaction, a new entity will emerge that is constituted by, but different from, the behavior of agents observed through Obs^1. If a property were non-emergent, it would mean that it could be generated from the structure observed at Obs^1.

Given the number of heterogeneous agents in MABMs (for instance, 1,000 households and 100 firms in Lengnick (2013), all with some heterogeneity), the inductive steps that allow us to observe a new level can, in the case of emergent phenomena, only be taken through simulation. However, it is essential to understand that the different levels of structure in MABMs are all inducible from the same input. This also means that the structure observed at any level is not independent of the structure observed through Obs^1. Dawid and Gatti (2018) also comes to this conclusion by showing that any macro property can be derived through mathematical induction.

2.3 Validation at Multiple Levels

In what follows, I relate the concept of emergent properties to validation in practice. I will do so by looking at how the properties at the micro, meso, and macro levels are validated. As examples of validation in practice, I refer to Lengnick (2013) since this study is, to some extent, representative of the general validation practices in the MAMB literature.

In particular, I will assess how validation as observed in Lengnick (2013) relates to the micro, meso, and macro levels. As stated previously, the distinction between the meso and macro levels is often not so easy to make in practice. I still think it is useful, however, to consider the meso level to better distinguish between different types of stylized facts. Importantly, the core of the analysis would remain unchanged if one were to consider an MABM as a two-level system.

Central to my discussion of the empirical validation conducted in Lengnick (2013) is the distinction between phenomenological input and output validation criteria. A validation criterion is a requirement that has to be met in order to consider the model valid. Phenomenological criteria are requirements derived from empirically observed phenomena. The distinction between input and output here refers to whether the implications of model assumptions are taken into account.

The Micro Level

Agents represent the entities at the $n = 1$ level, the micro level. They are the lowest level entities of an MABM, implying that they are the model input. They have properties that directly result from the assumptions made by the modeler. The validation of these properties can be labeled as input validation. In Figure 2.1, we can see a schematic representation of validation at the micro level. In this representation, the blue nodes can be seen as the agents E_i^1, and the edges as their interactions Int_{ij}^1. We are looking at the system through the Obs^1 lens; this means that we observe both the agents and their interactions. However, we do not consider the further structural implications of these interactions when addressing the system through Obs^1. P^1 are the properties observed through Obs^1, that represent properties of the individual agents. Input validation assesses whether these properties are in line with what is empirically observed.

What are some general characteristics of input validation in MABMs? The assump-

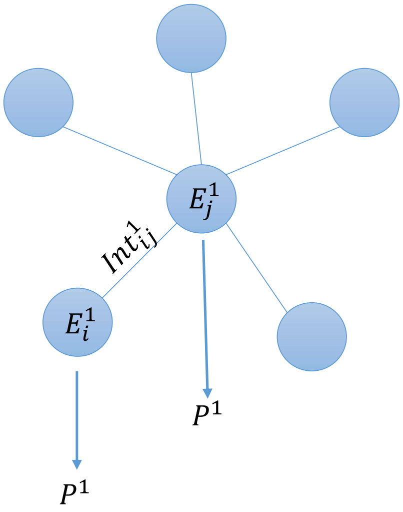


Figure 2.1: Validation at the micro level

tions at the micro level are considered to be of particular importance to MABM practitioners. MABM studies often start with the critique that DSGE models are not based on realistic assumptions. In turn, MABM modelers pride themselves on modeling from more realistic assumptions. But what does it mean, in the case of MABMs, to have more realistic assumptions? MABM practitioners state that one of the crucial differences between DSGE models and MABMs is in the realism of the assumptions that determine agent behavior. The following quote is one of the essential differences between neoclassical economics (in which DSGE are one of the main tools) and agent-based computational economics (ACE):

> *ACE can be seen as a substitute to standard neoclassical approaches to economics that tries to build more reasonable models based on reality to better address its behavior, a new approach that rejects the idea that models can be built using false assumptions and trying instead to explore models based on assumptions more in line with what we know about how real-world agents behave and interact.* (Gatti et al., 2018)

Hence, MABMs are an attempt to move from the homo oeconomicus toward a more empirically validated agent. In most MABM studies, therefore, specific attention will be paid to validation at the micro level, in which the assumptions regarding the behavior of agents are compared to outcomes of economics experiments or other insights from the psychology or strategy literature. Most often, these rules will imply that agents have a certain form of limited knowledge regarding the economic system when compared to the rational expectations approach of DSGE. This is in line with the concepts of procedural rationality (Simon, 1976) and heuristics (Gigerenzer & Todd, 1999). In the schematic overview, properties P^1 at the micro level reflect such behavioral rules.

Turning to how agent validation takes place in Lengnick (2013), we can observe a similar means of agent validation. For example, to describe the consumption behavior of a household h, the following equation is used:

$$c_h = \left(\frac{{m_h}^{\alpha}}{P_h}\right), \tag{2.3}$$

Where c_h is the individual household consumption, m_h are the monetary holdings of the household, and P_h is the average price of the producers from which the households buys. Here, $0 < \alpha < 1$ (parameter) such that the relative share of income consumed decreases when the household's monetary holdings increase. Importantly, the functional form of this equation is defended not on the basis of homo oeconomicus theory, but rather by citing an empirical study, namely Souleles (1999). For other domains of household and firm behavior, similar sources are cited that are either empirical studies or theories that are directly supported by micro data. It is useful to note that in DSGE models, such phenomenological input criteria are generally not present or addressed explicitly. Rather, the criteria at the agent level are theoretical. For example, consumers are assumed to optimize their utility based on their level of consumption and leisure. Such assumptions are not based on micro-economic evidence but stem from what is usually referred to as neoclassical theory.

The Meso Level

Let us now consider how properties at levels higher than the micro level relate to validation as we observe it in practice. In addition to validation at the micro level $n = 1$, we observe the validation of cross-sectional properties. The use of cross-sectional properties can best be characterized as validation at the meso level or $n = 2$, we can label these properties as P^2 of an entity E_i^2. E_i^2 arises from considering the structural implications of the first-order interactions between, for example, firm agents and their consumers. The network of firms and their consumers can be seen through the Obs^2 lens as a new entity. As P^2 requires us to consider the implications of assumptions, we can label them as phenomenological output criteria. An example of a property of these entities is the distribution of firm size, as these entities will vary in terms of how many links they have with consumers. These distributions can then be compared to empirical data to fulfill the phenomenological validation criteria.

Figure 2.2, presents a schematic overview of this. The second-level entities are given by the units represented as the large circles drawn around the network of squares and circles. These could represent a network of two groups of consumers and a firm. In Lengnick (2013), we observe such a use of cross-sectional properties in validation. For example, model data on the distribution of firm size are used. The model data follow a right-skewed distribution – more specifically, a power law. Practically, this means that the model data contain a large group of small firms and a small group of very large firms. In Lengnick (2013), these model data characteristics are compared to the empirical data on firm size, and it is concluded that both are distributed in a similar way. In some cases, this size distribution conforms to a Yule distribution, which is a key subject in Chapter 4. In addition, Lengnick (2013) uses model data regarding the distribution of the price changes for validation.

Validation of these distributions can be seen as validation of the meso level. This is because, especially in the case of firm size, these distributions are the product of interactions between agents rather than the product of interactions between networks of agents. Firm size is determined by its consumer network, and the distribution of these consumer networks can thus be seen as a property of E^2 as observed through Obs^2.

The Macro Level

Finally, we can look at the properties at the macro level E^3, with properties P^3. These properties can also be labeled as phenomenological output criteria because they require us to take into account the implication of the model input. These usually consist of patterns of aggregate variables over time or relationships between macroeconomic variables. In Figure 2.2, we can see a schematic overview of validation at the micro, meso, and macro levels. The rectangle around the groups of networks represents the model as a whole. Properties at the macro level are different from those at the meso level, because for the macro level, we account for the structural consequences of the interactions Int^2 between E^2, which yields a new entity E^3. If we consider how validation in practice is conducted at this level in Lengnick (2013), we observe the use of empirical regularities in the form of relationships be-

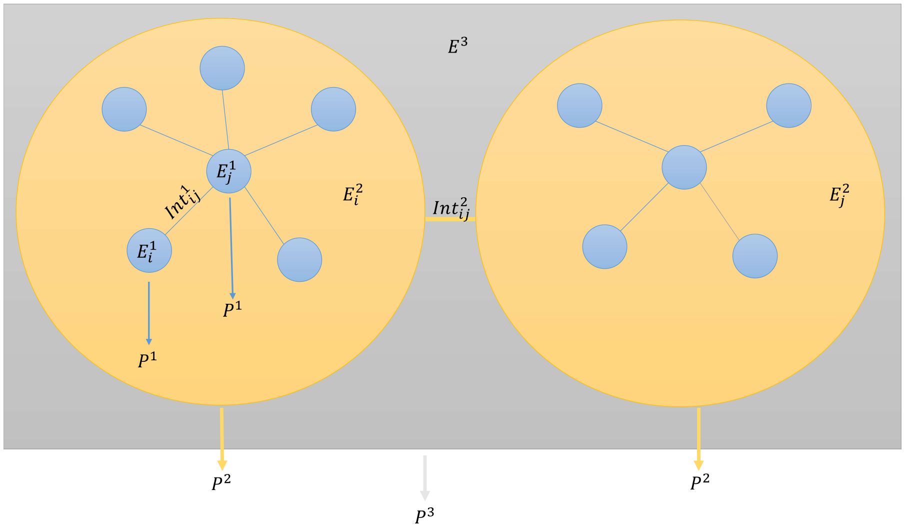


Figure 2.2: Validation at different levels

tween macro variables. Examples include the relationship between price level and employment, better known as a Philips curve, and the relationship between vacancies and unemployment, a Beveridge curve. For these relationships, the model data are compared to empirical data to fulfill the phenomenological criteria. In addition to examining the relationships between macro variables, we can also consider the characteristics of macro variables over time. An example of this is the unemployment level over time. In Lengnick (2013), the model unemployment data alternate between low unemployment and deep downturns. This is in line with what we observe empirically regarding the business cycle.

Again, whether a distinction exists between the meso and macro levels depends on whether the properties of the entities at the meso level can be obtained without considering the interactions between these meso entities, and whether the properties at the macro level require interaction between the meso entities in order for the model to generate them. A well-known example of this would be a skewed distribution of firm size that arises out of the interactions between firms and consumers and a business cycle that arises through the default of one large firm that affects other smaller firms due to decreases in the wage income of the defaulted firm's employees (Gabaix, 2011).

2.4 A Framework of Model Validation

To better understand the workings of the validation process of MABMs as observed in Lengnick (2013), we should first introduce some general concepts of model validation, such as model purpose and target and structural validation. In addition, I will introduce the classification of models presented in Boumans (2009), which is based on which validation tests are applicable .

2.4.1 Model Purpose and Types of Questions

Let us start with a brief discussion of model purpose, which will be further elaborated upon in Chapter 3. As we have discussed, models are constructed for a certain purpose. Often, this purpose is to answer a question. We can distinguish multiple types of questions that are typically associated with particular validation requirements. Here, I follow the question types formulated in Boumans (2009): why questions, how-much questions, and how's-that questions.

For why questions, the answer is an explanation. That is, it is an account of mechanisms that are to be interpreted as the mechanisms operating in the real world. An example of this type of research question is: Why did we observe stagflation during the Oil Crisis of the 1970's (Barsky & Kilian, 2000)?

For how-much questions, the answer is a measurement-a quantitative value. For this question type the answer does not need to contain an account of the mechanics by which the measurement was generated. An example in this case is: When will the next financial crisis occur (Aydin & Cavdar, 2015)?

The answer to how's-that questions does contain an account of the mechanisms; in contrast to why questions, this account does not have to be an explanation. This means that the account of mechanisms does not need to be interpreted as mechanisms operating in the real world. Rather, it is sufficient when the mechanisms behave *as if* they were in the real world in relevant domains. An example of this is: How do certain very ordinary economic principles lead maximizing individuals to choose consumption-production plans that display many of the characteristics commonly associated with business cycles (Long Jr & Plosser, 1983)?

2.4.2 Phenomenological Validation Criteria

In this section, we will discuss how the aforementioned question types determine which validation tests are relevant. To do so, requires a closer investigation of phenomenological validation criteria.

As we have discussed, phenomenological criteria come in the form of what is more generally known as empirical validation. This involves assessing whether certain characteristics of the model are in line with what is observed empirically. I will first discuss the various types of phenomenological validation criteria, as well as which tests are associated with these types. Thereafter, I will lay out a categorization of model types based on the types of phenomenological criteria that apply. These model types, in turn, are associated with particular types of questions that models are built to answer, as discussed in the previous section.

First, as introduced in Section 2.3, we can distinguish between phenomenological input criteria and phenomenological output criteria (see Gatti et al. (2018) for a similar categorization). Input criteria require the empirical assessment of the model assumptions without considering the implications of these assumptions when put together. That is, they are the initial settings of the model, before any analytical derivations or simulation have been performed. In the case of macroeconomic models this often comes in the form of assumptions at the agent level. We could, for example, seek to assess whether consumer saving behavior is in line with how subjects are observed to form expectations in laboratory experiments.

Phenomenological output criteria, on the other hand, require the assessment of the implications of the model assumptions when put together. This may include both the implications of all model assumptions or subsets of assumptions. Indeed, the implications of model assumptions can be studied in different ways. In some cases, the model is solved analytically; relationships within the model structure can be uncovered in this way and subsequently compared to relationships found in empirical data. In other cases, the behavior of the model as a whole is generated through computer simulations. This yields model output data, the characteristics of which can be compared to potential empirical counterparts.

The phenomenological output criteria can be distinguished further. Barlas (1996) distinguishes between three types of validation tests that can be understood as different types of phenomenological output criteria within the context of the framework presented here. The three types of tests are behavior pattern tests, direct structure tests, and indirect structure tests.

Behavior pattern tests can be understood as phenomenological output criteria directed toward the "major behavioral patterns of the real system" (Barlas, 1996). Which patterns are considered major should be understood in relation to the purpose of the model. For a model built to answer a question on how business cycles arise, for example, the cyclical patterns of business cycles would be considered major; they represent the most important patterns for a model to reproduce relative to its purpose. The major patterns in relation to the purpose of the model are what I will label as the model target. If the purpose of the model is to answer a question about a given phenomenon, the model target should be understood as empirical patterns closely associated with that phenomenon.

Next, let us discuss direct and indirect structure tests. Both types of tests can be understood as phenomenological output criteria directed toward the model structure. The model structure entails the mechanisms by which the model output is generated. Whether and the way in which the model structure is subjected to validation is dependent on the type of question that the model is constructed to answer. For instance, if the question is of the why type, an explanation is required, and this explanation is embedded in the model structure. This implies that, for these question types, the model structure is subject to validation. In contrast, how-much questions do not require an explanation. The answer to such a question is a quantitative measurement that is as accurate as possible; the mechanisms through which the model arrives at this measurement are not of interest. This implies that the model structure is not subject to validation.

Let us now discuss the difference between direct and indirect structure tests. Direct structure tests assess individual parts of the model structure, such as individual relationships between variables. For example, it may be presumed that a negative direct relationship exists between interest rates and the propensity to consume. A direct structure test would assess whether this direct relationship is, in fact, present in the model as such. Direct structure tests therefore assess whether the explanation provided by the model is in line with the description of the real-world structure (at least in the way that this is observed). I will also refer to this type of validation test as *direct structure validation*. To relate our discussion here to that of the elements from which models are constructed in Section 1.2, direct structure validation assesses whether the elements are injective representations.

Phenomenological input criteria can also be seen as a type of direct structure validation. This is because the model assumptions are input to the model structure, and the validation of these assumptions occurs without considering their relation to the other assumptions or their implications for the rest of the model.

The second type are structure-oriented behavioral tests. These tests assess the model structure, but they do so by comparing a broad range of characteristics of the model output data with empirical counterparts. The idea is that if the model behaves in line with observed facts in a sufficient number of dimensions, it signals that the model behaves at least as-if it were the real-world structure. If a model is built to provide an understanding of the mechanisms that could generate the business cycle, for instance, the fact that it is also able to reproduce the Philips curve is seen as a sign that the model mechanisms behave akin to the real-world structure. I will refer to this type of validation test as *indirect structure validation*.

Figure 2.3 presents an overview of all the different types of phenomenological validation criteria discussed thus far. The first distinction is between validation based on model input and validation based on model output. Output validation can be split into target and structural validation. Structural validation, in turn, comprises direct and indirect structural validation.

2.4.3 Black, White, and Grey-Box Models

Based on which forms of empirical validation apply and the types of questions that models are constructed to answer, several authors have put forward a characterization of model types. Barlas (1996) distinguishes between two types of models: white-box and black-box models. Boumans (1999) extends this classification to include grey-box models; as we will see, these are of particular relevance for models in economics.

Let us now proceed with a discussion of the model types. For each model type, a question is typically associated with it, as well as forms of empirical validation that apply. To begin, target validation applies to all three white-box, black-box, and grey-box models, because the purpose of the model is most often a question about a

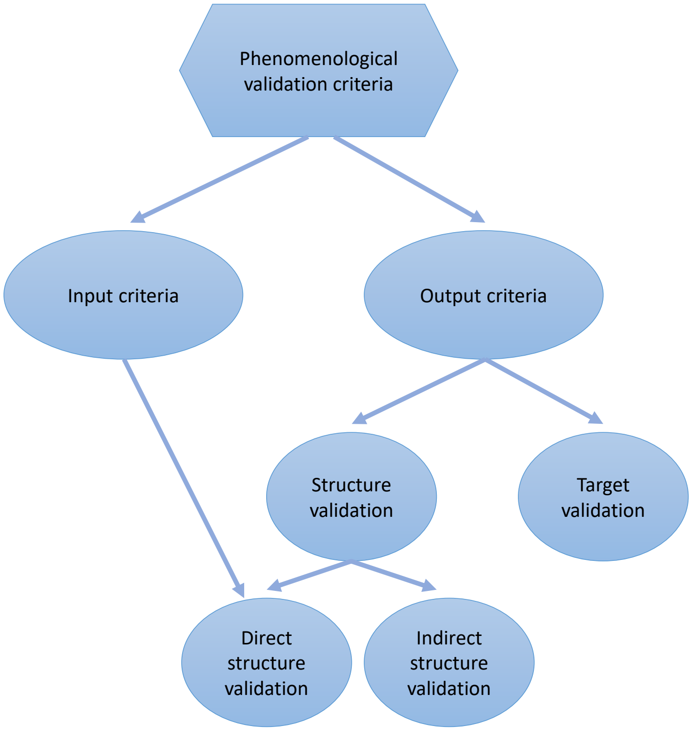

Phenomenological validation criteria
Input criteria
Output criteria
Structure validation
Target validation
Direct structure validation
Indirect structure validation

Figure 2.3: Validation criteria

phenomenon. This phenomenon is what we have defined as the model target. Irrespective of whether the answer to the question requires an account of the mechanisms involved, the model should at least be able to reproduce its main target.

Target validation is thus not a factor that allows us to distinguish between the various model types and types of questions. Rather, the types of structure validation that apply are what enables this categorization. Recall that the types of structural validation are indirect and direct structure validation.

White-box models have a structure that is typically associated with why questions. Both types of structure validation apply here. White-box models aim to be an injective representation of the structure of the system under investigation. If the system is large and complex, such as the macroeconomy, the white-box model structure is typically large and complex as well. In economics, given their complexity, white-box models are most often constructed in terms of macroeconomic variables. This is because aggregation is typically associated with a reduction in the number of variables. Take inflation, for example: rather than accounting for price changes of individual products in the economy, we can look at the aggregated inflation level. A primary example of white-box models in macroeconomics are simultaneous equation models (SEMs), such as those produced by the Cowles Commission. Another example is the Brookings model of the United States (Dusenberry, Fromm, Klein, & Kuh, 1965), which consists of over 400 equations.

Black-box models are typically associated with how-much questions. Structural validation does not apply here. Black-box models do not aim to represent the real-world structure in any way, nor is their goal to provide some form of understanding; they are evaluated solely based on their predictive accuracy. We can roughly distinguish between two sub-types of black-box models. The first are models whose structures are highly complex. Think, for example, of models generated by machine learning algorithms such as neural networks. The second sub-type is more simple but relies heavily on stochastic and/or temporal lags, which do not have a structural interpretation. In macroeconomics, black-box models are most often associated with this second sub-type and fall under the umbrella of time-series econometrics. An example are vector autoregressive (VAR) models, which are constructed from lagged macroeconomic variables and stochastic shocks.

Third, grey-box models are associated with how's-that questions. Only structure-oriented behavior tests or indirect structure validation apply here. The structure of

	Direct structure validation	Indirect structure validation	Target validation	Question type
White box	✓	✓	✓	Why
Grey box		✓	✓	How's that
Black box			✓	How much

Table 2.1: Model categorization based on phenomenological criteria and question type

the model is intended to be a non-injective representation. As Boumans (2009) points out, grey-box models typically have a "modular" structure. Following Simon (1962), such models are built from several, in some sense, autonomous parts or sub-models. The reason for modeling in this way is that it may be a daunting task to model all the structural elements directly that are at work in a larger system. Instead, we can seek to partition the system into smaller subsystems (or modules). If we put these modules together, we have a model of the larger system. These modules may interact in complex ways that make it difficult to validate the relationships between model entities individually, which leaves us with indirect structure validation. Typical examples of grey-box models in macroeconomics include DSGE models, into which modules such as Calvo pricing and intertemporal utility optimization are integrated. As I will argue in this chapter, agent-based models can also be seen as a particular type of grey-box model in which the modules are the agents.

Table 2.1 provides an overview of the categorization made in the previous sections. The types of models can thus be associated with the types of questions and the types of structure validation that apply.

2.4.4 Validation as Reduction of Underdetermination

Thus far, we have discussed several types of phenomenological criteria and how we can categorize models based on which of these apply. Now let us move onto how the fulfillment of these criteria enhances the perceived ability of a model to fulfill its purpose from an epistemological point of view.

To see how phenomenological validation criteria can enhance perceived model correctness, let us first introduce the concept of the model domain. This domain is information embedded in the model that has the potential to be empirically falsified.

That is, it is the information within the model that can be interpreted as an empirical claim.

This domain is determined through information embedded by the model's input and output that is empirically falsifiable. For instance, consider a model that produces a cyclical pattern of a variable labeled as *output gap* through model input that includes a negative relation between a variable labeled as *unemployment* and one labeled as *inflation*. Two distinct empirically falsifiable pieces of information are embedded in the model input and output. The first are the cyclical patterns in the variable labeled as output gap. If these cyclical patterns do not correspond to those that are empirically observed, the model's answer is empirically false in this respect. The second is the negative relation between a variable labeled as *unemployment* and one labeled as *inflation*, which is empirically false if it does not correspond to the relationship between unemployment and inflation as it is observed empirically. The domain, in this case, are the cyclical patterns in the output gap produced by the model and the negative relationship between inflation and unemployment.

Note that not all information within the model domain is necessarily relevant for the purpose of the model. Information that is relevant is defined as information that enhances the validity of the model given its purpose. To clarify, some elements in the domain are determined through information that is embedded in the model target, while others are determined through information that is embedded in the model structure. To take the aforementioned example again, the first piece of information – the cyclical patterns – represents the model target. The input of the negative relation between inflation and unemployment represents the model structure. In the previous sections, we have established that whether or not the model structure is subject to empirical validation depends on the model purpose. For how much-questions, for instance, the model structure is not subject to validation. To illustrate, let us say that the negative relationship between unemployment and inflation is in fact not observed empirically. If the purpose of the aforementioned model is to measure the duration of business cycles and not to explain them, the fact that part of the information in the model is empirically false does not negatively affect the validity of the model. To connect the notions of domain and relevance, therefore, we can say that the notion of relevance determines which subsets within the model domain enhance the validity of the model.

Importantly, any set of available empirical data is always underdetermined by the information in the model domain. That is, in principle, there will always be multiple ways in which a model can reproduce any set of data (Stanford, 2009). We can say that for a given set of data that does not falsify the information embedded in model A, we can always come up with a model B, the domain of which is also not falsified by the same set of empirical data. As a result, the goal of phenomenological validation criteria is not to definitively establish that the model domain is empirically true but to make it less likely that it is false. In this sense, phenomenological validation criteria serve to reduce underdetermination.

Given this epistemic value of phenomenological validation criteria, note that the more empirical data characteristics are involved in the validation process, the more underdetermination can be reduced as long as these facts are within the domain of the model's answer. This is in line with the view of some modeling practitioners that models that are able to reproduce a larger number of empirical facts are more epistemically valued.

2.5 Structure and Target Validation in MABMs

In Section 2.3, we discussed how phenomenological criteria at different levels apply to MABMs. We distinguished between phenomenological input criteria at level $n = 1$ and phenomenological output criteria at higher-order levels. To situate MABMs within the white-, black-, and grey-box framework presented in the previous section, it is important to further distinguish these phenomenological criteria in accordance with Figure 2.3 in terms of target validation and the various types of structural validation.

The properties observed at Obs^1 (i.e., properties of agent behavior) are used to fulfill phenomenological input criteria, which is a type of direct structure validation. Agent behavior is modeled based on empirical evidence related to agents, as shown in Section 2.3.

This agent behavior, as observed through Obs^1, also constitutes the input for the structure observed through Obs^2, the system of interacting agents. This structure, contrary to the structure observed through Obs^1, is usually not observed by the modeler directly. The reason for this is that MABMs can, in practice, only be analyzed through (numerical) simulation due to the large number of heterogeneous agents engaging in non-linear interactions. The structural elements that are not observed directly can only be validated indirectly through, for example, assessing the reproduction of both the cross-sectional and the time-series levels. In the previous section, I have shown examples of such validation. Importantly, both the assessment of the reproduction of properties at the meso level as well as the macro level can be seen as indirect structure validation as long as these properties are different from the model target. The model target in the example of the previous section are the observed business cycle dynamics. If the property coincides with the model target, this would be characterized as target validation and not structure validation.

Looking back at the classification by Boumans (2009) in Table 2.1, we observe that MABMs are thus validated as grey-box models. This is because MABMs have a modular structure in which the modules are the agents. Together, the structure that these models create is validated indirectly. Specific to MABMs, these modules are assessed through direct structure validation in the form of input validation, meaning that we can view them as white-box sub-modules. Within this classification, MABMs can be seen as a grey-box model built from white-box modules.

Now, how does this combination of direct structure validation and indirect structure validation enhance model validity? If the relationship between multiple levels is merely one of inductive steps, does this mean validation at higher levels is superfluous? To understand why the answer is no, it is useful to think about this in terms of reducing underdetermination, as explained above. When modeling agent behavior, the behavior of the agent is constrained by empirical data through direct comparison of agent behavior with empirical evidence related to agents. However, this does not mean that choices regarding agent behavior do not have to be made. In particular, two sources of degrees of freedom are available.

First, modeling is always a matter of simplification and isolation. The modeler must decide which elements of agent behavior are significant in relation to the model target and which elements can be omitted. Second, underdetermination can only be reduced to some degree, since there will always be degrees of freedom and several competing models of agent behavior, even if we apply strict empirical testing. The above facts are important to state, since they imply that indirect testing of the structure at Obs^2 will still enhance the validity of the model's structure because it further reduces the number of possible specifications of agent behavior that are in line with empirical reality. The modeler makes choices regarding agent behavior when modeling the agent; the innovation that MABMs have brought is that these choices are constrained by the empirical data regarding the behavior of the individual agent.

On the other hand, the choices that the modeler makes also determine the structure at higher levels of observation. By simulating the model, the implications of the choices made at the level observed through Obs^1 for the level observed through Obs^2 can be uncovered and indirectly validated. This further constrains the choices that the modeler can make regarding the model input. The role of emergence in this process is thus merely one of unveiling the implications of choices at one level for the next level. The structure at higher levels can only be uncovered through some inductive process, which in the case of MAMBs is numerical simulation. The simulation methodology implies that this higher-level structure, as implied by the lower-level structure, can only be validated indirectly through output characteristics.

If no emergence relationship existed between two levels, there would be no veiled implications for the choices of one level for the next. Validation at these levels in such a case would therefore require the same data, which would mean that validating at multiple levels would not help to reduce underdetermination.

2.6 Conclusion

In summary, to understand the validation practices conducted in MABMs, I have introduced several frameworks.

First, I have looked at how models of complex systems with emergent properties are structured. Complex systems can be seen as systems with different levels of interacting entities. Properties observed at one level that are not observed as properties at a lower structure can be defined as emergent properties. In relation to this, I have looked at how models of complex systems with emergent properties are structured. Complex systems can be seen as systems with different levels of interacting entities. Properties observed at one level that are not observed as properties at a lower structure can be defined as emergent properties. MABMs can also be understood as systems with multiple levels of entities: the micro level, the meso level, and the macro level.

Second is a classification of three types of models based on how these models are validated and which questions they are constructed to answer. We can distinguish between black-box models in which only target validation applies, white-box models in which both the target and the structure are subject to validation in a direct sense, and grey-box models in which the structure is subject to validation in an indirect sense. This classification helps us understand how we should consider the purpose of validation tests used in practice.

The combined insights of these three elements allow us to formulate a validation methodology for MABMs. First, I have shown that in practice, each of the levels (micro, meso, and macro) of MABMs are validated using different tests. The micro level is the input to the model. Specific to MABMs, this input is subject to validation. This means that the agent behavior should be similar to the agent behavior observed empirically in relevant domains. The meso level is validated by comparing model output to the distributions of aggregate variables. Finally, properties at the macro level are validated by comparing model output to the behavior correlations between aggregate variables over time.

Next, we examined how we should understand MABM validation in the context of structure and target validation. Again, the starting point is the layered structure of MABMs. This implies that the structure at one level provides input to the structure at the next level. The structures embedded in the agent behavior are validated by comparing their behavior with the empirical behavior of agents directly. These structural elements then form input for the structural elements at higher levels, which cannot be observed at the micro level since they are emergent properties. Rather, they arise out of agent interaction through an inductive process. Due to the complexity of MABMs, this process usually comes in the form of numerical simulation. This implies that the structural elements at higher levels can only be validated indirectly. This combination of direct and indirect assessment of the model structure means that MABMs can best be described as grey-box models built from white-boxes.

In conclusion, this chapter serves as a step toward a validation methodology and provides a more systematic understanding of current MABM validation practices. We have seen that a proper assessment of validation practices in MABMs requires an understanding of both model validation and complex systems with emergent properties. The validation of MABMs cannot be compared one-to-one with DSGEs or other types of macroeconomic models. Crucially, MABMs have unique structures, which has led to specific forms of validation practices that are new to macroeconomics.

In Section 6.4 of the conclusion of this dissertation, I will provide a more in-depth discussion of realist and instrumentalist views on modeling as applied to MABMs, which builds on the discussion of this chapter.

Chapter 3

DSGE Models and the Hybrid Model Critique

3.1 Introduction

Since the 2008 financial crisis, DSGE models have drawn criticism on several fronts. Most of the criticism has targeted unrealistic model assumptions with respect to behavior at the agent level, as well as dimensions of the real economy that were missing in the models altogether (Stiglitz, 2018; Christiano, Eichenbaum, & Trabandt, 2018). For the former, rational expectations were a common target; it was seen as unrealistic that agents would be endowed with so many cognitive capabilities. For the latter, the omission of financial frictions was seen as a primary reason why DSGE models were not able to be of use during the 2008 financial crisis. DSGE modelers have responded to these criticisms by incorporating insights from behavioral economics in expectation formation, such as heuristics and learning; in addition, they now often incorporate financial frictions as well.

In my view, however, another type of criticism remains insufficiently discussed. It is laid out in, for example, De Grauwe (2012) and Chari et al. (2009)and concerns the way in which contemporary DSGE models are constructed and, subsequently, empirically validated. First popularized in (Smets & Wouters, 2003), DSGE models are generally empirically validated by estimating the model on time-series data and then comparing the model output to its empirical counterparts.

Such models (Smets & Wouters, 2003) are able to reproduce the behavior of the real economy after shocks well. However, the problem is that DSGE models can only reproduce this behavior after the supplementation of a large number of stochastic terms (and lags of stochastic terms). To put it differently, the model is supplemented with a stochastic process that is structurally different from the core model.

According to Chari et al. (2009) this stochastic supplementation to the model lacks a representational interpretation and is thus primarily a tool to ensure that the model can be successfully estimated and validated. The term representational here, refers to the notion that the model is to be interpreted as representational of the economic structure. Representational models aim to provide an account of the economic mechanisms, which is seen as necessary to guide economic policy. For instance, consider a model supplemented with a normally distributed shock to labor demand. A sudden drop in labor demand can be understood through many economic mechanisms. The shock to labor demand by itself, however, does not make explicit reference to any of them, it is a black box. Chari et al. (2009) states that this is problematic if the purpose of the model is to guide economic policy because it does not identify the mechanisms through which policy may work.

As we have discussed in Section 1.2, economic models are not to be interpreted as purely representational; rather, they also rely on artifactual elements. Therefore, when I refer to representational and non-representational model structures in this chapter, I do not imply that one does not rely on artifactual elements while the other does. A representational structure is one that explicitly describes an economic mechanism. A non-representational structure is most often constituted by stochastic or lagged elements that are a *stand-in* for various possible economic mechanisms. Note that a representational model structure can include both injective and non-injective representations, as defined in Section 1.2. Given that DSGE models are primarily validated using indirect structure validation, their structure can be characterized as more in line with non-injective representation (see Section 2.4.2for a discussion on validation and representation).

In a different formulation of the criticism, De Grauwe (2012) states that this stochastic supplementation produces models that are in line with what is observed even if the model structure does not map onto the real-world structure in a useful way. The stochastic process is able to capture most departures from the core (typically new Keynesian) equilibrium structure of the model, but it prevents new empirical insights from being translated to alterations in the model structures. Furthermore, given the non-representational nature of the stochastic terms, Stiglitz (2018) argues that the approach is susceptible to the Lucas critique.

A way to view this structure of DSGE models in this context is as a hybrid model – with a representational core and a non-representational , stochastic periphery. This description may remind the reader of the description of scientific research programs in Lakatos (1976). In this description, a scientific research program consists of core assumptions and a "protective belt" of non-core assumptions. New insights or data that do not agree with a scientific research program are not translated into alterations of the core assumptions but are usually captured by expanding the protective belt. Indeed, De Grauwe (2012) formulation can, in my view, be seen as a translation of this description to the context of DSGE models.

In short, De Grauwe (2012), Chari et al. (2009) and Stiglitz (2018) formulate a critique of DSGE models that rely on non-representational stochastic processes that are used to match the model to the data. To be used for policy, DSGE models require a representational interpretation. This implies that the model structure is subject to validation, which comes in the form of matching modeled data to empirical data. The fact that the model relies on non-representational elements to accomplish this weakens the model's validity. From now on, I will refer to this critique as the hybrid model critique.

The hybrid model structure is not tied to the original formulation of DSGE models, although it was introduced to make maximum likelihood estimation feasible. Before this estimation approach was popularized in Smets and Wouters (2003), it competed with calibration as an approach to the parametrization of DSGE models. The calibration approach was most prominently laid out in Kydland and Prescott (1996) and centers around the idea of parametrization by using relevant stylized facts or microeconomic data. The approach generally incorporates a much smaller number of stochastic terms compared to estimated DSGE models. I will elaborate on these issues later in this chapter.

However, present DSGE models have converged toward the estimation approach. They are taken to the data using, in essence, the same approach as Smets and Wouters (2003) (Fernández-Villaverde & Guerrón-Quintana, 2021). Given the relevance of DSGE models in current macroeconomic science and policy, an in-depth analysis of the hybrid model critique is required.

The goal of this chapter is to gain a more in-depth understanding of the hybrid model critique of DSGE models. This includes an analysis of how the structure of

DSGE models was developed into hybrid form, as well as an assessment of the validity of the hybrid model critique. Central concepts in this analysis are model scope, invariance, and validation. Model scope refers to the domain and type of overlap between the real world and the model. Invariance is the extent to which the model structure remains correct given changes in relevant context variables, such as time, place, or policy regime. We have defined validation as the assessment of a model's ability to fulfill its purpose (see Section 1.5.2. These concepts will allow us to understand how the shift from the calibration to the estimation approach implied a shift toward a hybrid model structure, as well as providing a more systematic understanding of the hybrid model critique.

This chapter builds on the idea that different types of models are relevant at different stages in bridging the gap between theory and data. Suppes (1966) is a well-known study in which a hierarchy of models is introduced. On top of this hierarchy are models that follow directly from a certain theory. Suppes (1966) provides an example of linear reinforcement learning in psychology. This theory consists of an axiomatic structure that can be translated into a mathematical model from which implications can be deduced. However, it is shown that such a theoretical mathematical model does not apply to data generated by relevant physiological experiments in a straightforward sense. Concepts applied in theoretical models, such as infinity and continuity, are generally not applicable to the context of empirical data, which are both finite and discrete.

To bridge the gap between the theoretical model and experimental data, we have to move down in the hierarchy of models to *models of data*. These require supplementation of the structure of the theoretical model. For instance, statistical concepts like goodness-of-fit and an error term need to be introduced. These concepts build on fundamental statistical notions, including the existence of a data-generating process.

The addition of a theoretical model to a data model is relevant in any scientific practice that involves models and data. That said, Suppes (1966) describes that some scientific practices are more strongly oriented toward theory models, whereas others are more oriented toward data models. In the former practice, more supplementation is required when a model is to be applied to data, and less so in the latter. In such practices, the gap between the theory model and the data model is smaller.

Let me now relate the work by Suppes (1966) to our discussion of DSGE models. DSGE models start from a well-articulated theoretical model. Yet the path from this theoretical model to the data is not straightforward and requires the supplementation of various elements. As we will see, the calibration and estimation approaches are different methods for navigating this path. The calibration approach does so mainly by limiting and particularizing the data, while the estimation approach does so mainly through the above-described hybridization of the model structure.

This chapter builds on the concepts of representation introduced in Section 1.2 and on model purpose and phenomenological validation criteria in Section 2.4 of the previous chapters. In addition, this chapter will start by introducing the concepts of model scope, invariance, and theoretical and technical validation criteria, which are necessary to understand the rest of the chapter. Next, we proceed to analyze both the calibration and the estimation approaches to DSGE models. Finally, we will analyze the hybrid model critique itself.

3.2 Validation Criteria, Invariance, and Model Scope

3.2.1 Validation Criteria

To understand how DGSE models evolved from a calibration approach to an estimation approach, we need to gain a broader understanding of the validation criteria that are relevant in the model construction process. In addition to the phenomenological criteria discussed in Section 2.4.2, what I label as theoretical and technical validation criteria are relevant to understanding the evolution of DSGE models.

Let us first discuss theoretical validation criteria. Theoretical criteria constitute the assessment of whether a model is in line with established relevant theory. Let us first establish what is meant by theory, since there have been multiple ways in which theory has been defined in relation to economic modeling. For example, Lucas (1980) broadly defines theory as the set of instructions to build the model economy. This is not the definition of theory that I will use throughout this dissertation.

The way in which I use the notion of theoretical validation criteria in this dissertation is as existing conceptualizations of the mechanisms associated with the real-world structure. New Keynesian economic theory, for example, entails price stickiness. A macroeconomic model that is to be in line with new Keynesian theory should be able to generate price sticky behavior if it is to meet such a theoretical criterion.

Next come technical validation criteria. As discussed in Section 1.2, a useful metaphor for models is to view them as mediating instruments of investigation. Models are tools that help us answer particular scientific questions. Because these questions are about phenomena observed in the world, models incorporate, to some extent, representations of the world. The purpose of the theoretical and phenomenological criteria that we have discussed so far is to assess these representational aspects of models. However, this is not enough to enable the model to answer the question correctly. The ability of models to answer questions also relies on a technical dimension, one which is assessed based on technical validation criteria. In the context of macroeconomic modeling, we can distinguish between mathematical and statistical criteria.

In order for the model to provide answers to a question, it should work mathematically. This could refer to the analytical solvability of the model, which will play a role in the case of Chapter 4. In this case, if the model could not have been solved analytically, it would be impossible to consider its implications. The mechanisms of such a model could not be studied, and its output could not be compared to empirical data. Mathematical criteria, therefore, also include the related notion of analytical tractability. If the model purpose is to provide an explanation, it is implied that such an explanation enhances understanding. In such cases, limits to the mathematical complexity of the model are unavoidable.

Statistical criteria require constructing the model in such a way that a certain statistical methodology can be applied. One of the reasons why statistical methods are employed when taking the model to data is to ensure a degree of invariance. In what follows, I will make this concept more concrete.

The main way in which statistical methods are applied in the context of modeling is parametrization of the model – that is, to assign a quantitative value to the model parameters. Parametrization is a necessary step in most modeling exercises because it provides us with information regarding the “size” and the “sign” of the relationships within the model structure. If we are interested in the effect of an increase in interest rates on consumption propensity, we would want to know whether the effect is positive or negative and how much the effect will be. Additionally, parametrization is necessary before any model outcome can be compared to empirical data.

Parametrization techniques are statistical in nature, meaning that they make use of empirical data and certain assumptions about the data-generating process of the

	Description
Theoretical criteria	Require the model to be in line with existing theoretical concepts
Phenomenological criteria	Require various elements of the model's input, output, target, or structure to be in line with what is observed empirically
Technical criteria	Require the model to have favorable mathematical or statistical properties

Table 3.1: Summary of validation criteria

model and the real world. Importantly, the model structure has specific requirements before specific statistical methods can be applied.

The main reason for employed statistical methods in the parametrization process is to ensure some degree of invariance. For the model to provide answers to questions that remain stable over time, place, or policy regimes, the model parameters must remain stable as well. One example, which we will discuss further later in this chapter, is the calibration methodology that started with the Lucas critique (Lucas, 1976). In this methodology, models should be constructed from "taste and technology" parameters. The idea is that such "deep" parameters do not change as a result of policy interventions. The parameters are then usually indirectly calibrated using stylized facts, which have been shown to be relatively invariant over time.

Statistical methods are also applied to provide an "objective" measure of model performance. This is crucial when it is of interest to know whether a particular innovation in the model structure represents an improvement in terms of pre-existing model structures. An example is the likelihood score, which requires the model to be parametrized using maximum likelihood methods.

Table 3.1 provides an overview of the three main validation criteria discussed in this dissertation. Together, they should be seen as the main drivers of the choices made during the model construction process. The reason for this is that the fulfillment of validation criteria is generally built into the model (Boumans, 1999). These validation criteria are known to the model builder during the construction process, and model construction is often a back-and-forth process until the criteria are fulfilled.

3.2.2 Model Scope

The model scope refers to the domain and type of overlap between the real world and the model. The relationship between model scope and the discussed validation criteria is that the latter require a model of a certain scope. In other words, they require the model to overlap with the real world in certain ways to be fulfilled to a sufficient degree.

To introduce the concept of model scope, it is useful to introduce the idea of the real-world data-generating process (rwDGP) and the model data-generating process (mDGP; Windrum et al. (2007)). The rwDGP is to be considered a process instantiated by the real-world structure. Each variable that we observe can be thought of as being generated by a web of structural factors that determine its value. However, we cannot directly observe the rwDGP. Instead, we seek to gain a level of understanding of how the rwDGP works through the application of various scientific methods, including the construction of models. The mDGP generates data by running simulations with the model. The idea of scope thus centers on the relationship between the rwDGP and mDGP.

This relationship can be thought of as two-dimensional. The first dimension is the relationship expressed in terms of the degree of overlap in the relevant data generated by the rwDGP and the data generated by mDGP or, to put it differently, the degree to which the mDGP captures the observed data generated by the rwDGP.

The second dimension concerns the relationship in terms of the structures by which the rwDGP and mDGP generate data. The structure of the rwDGP can overlap with the structure of the mDGP while holding the overlap in terms of the first dimension constant in three main ways . First, there may be no overlap: the model structure and the real-world structure operate differently and have no epistemological relationship. That is, studying the model mechanisms yields no understanding of the rwDGP. The second case is where overlap is evident between the mechanisms in the mDGP and the structure of the rwDGP. That is to say, the mechanisms in the mDGP are injective representations of the actual structure of the rwDGP. In the third case, a different kind of overlap is found between the mechanisms. In this case, the mechanisms in the mDGP behave *as if* they were those in the rwDGP in terms of the wide range of outputs that the mDGP reproduces. This means that the mechanisms are not the same, but for the purpose at hand, they function well enough to be treated as if they were. The model structure is a non-injective representation of this wide range of outputs, and it can therefore be described as artifactual to a large degree. Let us refer hereafter to these two types of model scope as *output scope* for the first dimension of model scope and *structural scope* for the second dimension of model scope.

Note that the three types of structural scope that we have distinguished coincide with the three types of questions and phenomenological validation criteria in Section 2.4. We have established that different types of questions require different types of validation criteria to be fulfilled. This fulfillment, in turn, requires the right model scope. Target validation requires the right output scope, direct structure validation requires overlap as an injective representation, and indirect structure validation requires overlap as a non-injective representation. Furthermore, technical and theoretical validation criteria may also impose particular requirements on the model scope. This connection is important to keep in mind as we proceed to discuss the calibrated and estimated DSGE approaches.

Figure 3.1 presents a schematic overview of the notion of model scope. The rwDGP produces empirical data, and the mDGP produces model output data. The output scope concerns the overlap between these elements. The rwDGP and the mDGP embed a real-world structure and a model structure, respectively. The overlap between these elements determines the structural scope. Note that there is generally no complete overlap in terms of output: the mDGP will not generate all the observed empirical data generated by the rwDGP, nor can all of the simulated data be matched to the empirical data. This implies that empirical validation is always a matter of degree: How much overlap is there between the simulated and empirical data? In other words, it is an assessment of whether the output scope is sufficient.

3.3 How DSGE Models Evolved into the Hybrid Structure

Let us now investigate how DSGE models developed into their current hybrid form. As stated in the introduction, the current hybrid form of DSGE models is not necessarily tied to the origins of DSGE models. Rather, an explicit transition has taken place toward the hybrid model strategy (Fernández-Villaverde & Guerrón-Quintana, 2021). An investigation into the history of this transition is crucial, as it will provide us with a more in-depth understanding of the validity of the hybrid model critique. As I will argue, the change toward the hybrid model structure was necessary to en-

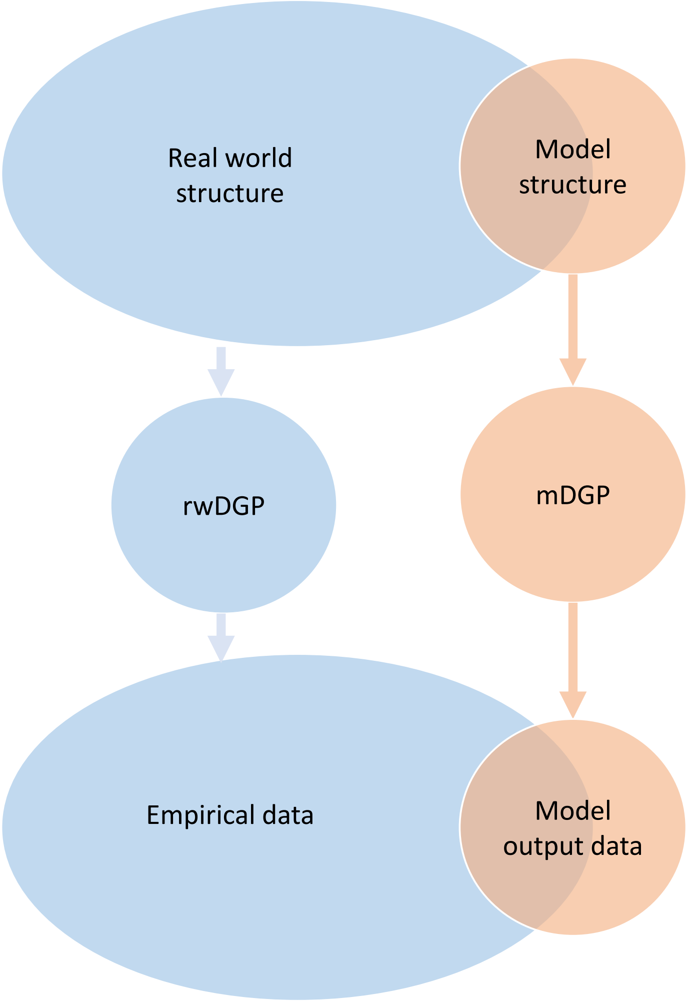


Figure 3.1: Model scope

able a shift in the parametrization approach of DSGE models – specifically from the calibration approach to the estimation approach.

3.3.1 The Calibration Approach

During the 1970s, Robert Lucas developed the foundations for what was later to become the DSGE approach. His approach centered on two main notions. The first were rational expectations, meaning that the expectations of agents are model-consistent. Agents are assumed to calculate the expected value of future-state variables using all the information present in the model. The second was an insistence that the model be constructed from invariant parameters. These are parameters that are likely to be stable over time and, moreover, across different policy regimes. This second foundation followed from the well-known Lucas critique (Lucas, 1976), in which he criticized macroeconomic simultaneous equation models for not being invariant to changes in policy regime.

One of the most well-known applications of the modeling strategy suggested by Lucas comes from Kydland and Prescott (1982). Their study also introduces the concept of calibration as a means of taking the model to the data. Kydland and Prescott (1996) discusses this same calibration approach on a more methodological level. Here, I label their methodology the calibration methodology. In what follows, I will review their methodology and connect it to the concepts introduced so far.

The calibration methodology in Kydland and Prescott (1996) is divided into five steps: pose a question, use well-tested theory, construct a model economy, calibrate the model economy, and run the experiment. Each of these steps forms part of what is labeled a computational experiment: answering a question by running simulations through a calibrated model.

The first step in this computational experiment is to pose a question. Answering this question is what is defined as the purpose of the model. In Section 2.4.1 of the previous chapter, we discussed various types of questions that models are typically constructed to answer: why questions, how-much questions, and how's-that questions. Given this context, we can ask to which types of questions the calibration methodology applies. In its application of macroeconomic models, Kydland and Prescott (1996) focuses its methodology on questions that "ask about the quantitative implications of theory for some phenomena." In other words, this is a how-much question. That being said, the calibrated DSGE methodology extends to how's-that questions, in which the model structure does provide an understanding of a phenomenon. Long Jr and Plosser (1983), for instance, uses a calibrated DSGE to provide an account of the mechanisms through which business cycles may arise.

Evidently, we can see that the calibration methodology starts from a purpose that can be characterized as specific. We have a clearly defined question about a particular phenomenon. This notion of specific purpose contrasts with what can be labeled as a broad purpose. This distinction will become important to understand the difference between the calibration and estimation approaches, as will be discussed in the next section.

A model constructed for a specific purpose also places specific requirements on the degree of overlap between the model and the real world (i.e., the model scope). This is because both the output scope and the structural scope are determined by the specific purpose.

The second step in the calibration methodology is to use well-tested theory: "a researcher needs a theory that has been tested through use and found to provide reliable answer to a class of questions" (Kydland & Prescott, 1996). What is interesting is what is meant by theory and what is meant by reliable.

By theory, Kydland and Prescott (1996) follows Lucas by stating that a theory is an explicit set of instructions for constructing a model. This use of the term theory therefore means something broader than just incorporating representations of theoretical notions into the model structure as discussed in Section 1.2, although it remains a vital part of it. As a prime example of this step, Kydland and Prescott (1996) mentions general equilibrium theory as being useful for answering business-cycle-related questions. This step thus presents a theoretical validation criterion: The model should be constructed in such a way that it is consistent with general equilibrium theory. Kydland and Prescott (1996) acknowledges that deeming which well-tested theory is appropriate is a function of the model purpose. As an example, they state that general equilibrium theory is not an appropriate choice for addressing the phenomenon of wealth inequalities between countries. The use of a particular theory thus provides a structure and output scope that is appropriate for some phenomena, but not for others.

In this context, reliable can be interpreted as similar to invariant. Invariance means that the model should be able to provide answers that remain stable over time and place and across different policy regimes. Only when this is the case can the answer provided be relied upon to, for example, guide policy. As briefly discussed above, since the Lucas critique (Lucas, 1976), one way to enhance the invariance of a model is to construct it from deep parameters. Most often, this has implied parameters at the microeconomic level, such as "tastes and technology" (Lucas, 1976), or the idea that such parameters are less affected by changes of policy regime. The construction from invariant parameters could be seen as a statistical validation criterion. As discussed in the previous section, invariant parametrization is one of the primary functions of statistical criteria.

The third step is the construction of the model economy. Crucial in this step is balancing the amount of detail in the model with its computational feasibility. It is stated that economists often have to work with a model that is much simpler and more abstract than what one would ideally prefer. The reason for this is that the complexity of the model is at odds with its intelligibility and computability. This means that the more complex a model is, the more difficult it becomes to both simulate the model in a technical sense and to interpret the outcomes of the simulation. It should be noted here that computational feasibility limitations have dramatically decreased since Kydland and Prescott (1996). The limitation of intelligibility, however, is still relevant if the purpose of the model is to provide understanding.

This third "construction step" is fully in line with the notion of mathematical validation criteria discussed above. Again, emphasis is placed on the relationship between the required model complexity and the model purpose. The question that the model is constructed to answer determines the required level of complexity of the model. There is an incentive to construct a model that contains the minimum level of complexity as long as it is able to provide a reliable answer to the question asked. In practice, this is often an exercise of balancing different validation criteria. More complexity may lead to a model that is better able to reproduce certain facts about phenomena, but this may come at the cost of analytical tractability, for example. Note here again the relationship between the validation criteria and the output scope. Fulfillment of mathematical validation criteria limits the degree of variation that can be reproduced by the model. This stresses the emphasis in the calibration approach of constructing a model for a specific question rather than the construction of more general models that can be applied to multiple distinct questions.

The fourth step is the actual calibration of the model. Calibration is a means of parametrizing the model, that is, to provide numerical values for the model parameters. In its approach, Kydland and Prescott (1996) draws heavily on the calibration of measurement instruments similar to, for example, a thermometer. The general idea is that parameter values are found either through direct calibration or indirect calibration (Windrum et al., 2007). Direct calibration is the collection of data from which a parameter value can be distilled in a direct sense. In the case of DSGE models, direct calibration involves the use of microeconomic data. A parameter of a consumer's propensity to save, for example, could be distilled from consumer survey data. Indirect calibration involves selecting the parameter values that enable the model to provide answers to questions for which we already know the answers. To put it differently, the parameter values are selected such that the model is able to reproduce facts about phenomena. In order for indirect calibration to yield invariant parameters, however, the facts should be stable over time and across policy regimes. If this is not the case, the parameter values would change depending on when and under which policy regime they were calibrated.

How does this calibration step relate to the concepts discussed so far? Note that the issue of calibration is also discussed in Section 2.1.1 in the previous chapter. First, successful indirect calibration is done through the reproduction of facts about phenomena, which is how I have defined the fulfillment of phenomenological validation criteria (see Section 2.4.2). Second, coupled with the technical validation criterion of constructing the model from invariant parameters, the calibration approach provides a technical means of quantifying those parameters. The emphasis on the stability of the facts about the phenomena to reproduce tells us that calibration is about more than the fulfillment of phenomenological criteria. More specifically, it is a technical validation criterion that enhances the model's ability to provide invariant answers to questions.

The further question is how we should select the facts about the phenomena to be used in the calibration exercise (in addition to the facts being invariant). Here, Kydland and Prescott (1996) again emphasizes that the facts should be relevant with respect to the model purpose. A model that is constructed to explain business cycles, for example, should be able to reproduce quantitative facts about the business cycle and the empirical reflections of the structural elements involved in the generation of business cycles. Calibration allows the model to be parametrized by relying solely on data characteristics that are strictly relevant to the purpose of the model. As such, a model constructed to explain business-cycle dynamics is generally calibrated through data characteristics associated with those dynamics. An example is the autocorrelation of the output gap in time-series data, where the model is parametrized such that the model reproduces autocorrelation in the output gap similar to the autocorrelation that is observed empirically. The output scope of the model in the calibration approach should therefore include those data characteristics that are relevant to the purpose of the model and used in its calibration.

The fifth step of the calibration approach is running the experiment. This concerns simulating the model over many runs and computing the probability distribution of key variables in the model. This allows the model to provide a measurement and thus answer the question for which the model was constructed.

To summarize, the calibration approach starts from a specific and limited purpose: providing a reliable (invariant) and purely quantitative answer to a "well posed question." When translated into validation criteria, we can see an emphasis on using particular economic theory and reproducing observed economic behavior. For these criteria, it holds that they are rather strictly related to the question that the model is constructed to answer. To put it differently, particular theoretical, technical, and phenomenological requirements are appropriate for a particular question. In addition, various technical criteria ensure that the model parameters are invariant; the model is constructed from representational microeconomic parameters (taste and technology). These parameter values are quantified through the calibration approach by matching model output to relevant and stable facts about phenomena. Given the limited purpose of the model, the theoretical, technical, and phenomenological criteria yield a model of limited scope, meaning that its domain of application is limited in terms of both structural and output scope. In addition, establishing invariance through deep parameters and calibration does not necessarily require an increase in model scope. This is in contrast to the estimation approach, which we will discuss in the next section.

3.3.2 The Estimation Approach

This subsection will discuss what I will label the estimation approach as applied to DSGE models. As with our discussion of the calibration approach, I will refer to the concepts of model purpose, invariance, model scope, and various types of validation criteria. This will allow us to not only understand why we have observed the shift from the calibration to the estimation approach, but also to offer insight into the utility of both approaches and a better understanding of the hybrid model critique.

The estimation approach was popularized in Smets and Wouters (2003), Christiano et al. (2005), and Smets and Wouters (2007) (among others). Fernández-Villaverde and Guerrón-Quintana (2021) provides a modern methodological discussion of this approach that I will use as a basis for my analysis.

To begin, let us discuss the starting point of the model construction process for estimated DSGE models: the model purpose. Here, I will refer to our discussion of model types in Section 2.4. To gauge whether a general notion of model purpose is present in the estimated DSGE methodology, I will review the discussion on model purpose in the main early estimated DSGE contributions. In Smets and Wouters (2003) and Smets and Wouters (2007), model purpose is described in the following way. First, there is the notion that the model is "for the euro area" (Smets & Wouters, 2003) or "for the US economy" (Smets & Wouters, 2007). Second, the model is used to "analyze the effects of various structural shocks in the euro-area," to "estimate the relative contribution of various shocks to empirical dynamics," and to "calculate the potential output level, real interest rate and gaps" (Smets & Wouters, 2003). In other words, the model is to be used to answer a range of questions about the macroeconomic system.

Just by looking at how these questions are formulated, one may say that they are all how-much questions, the answer to which is a measurement. Providing measurements and predictions are indeed part of the purpose of estimated DSGE models (Fernández-Villaverde & Guerrón-Quintana, 2021). In addition, however, if we observe how the model is actually used in Smets and Wouters (2003), providing an account of the model mechanisms that generate such measurements sits at the core of the analysis. The purpose of such accounts is to provide understanding; the questions that the model in Smets and Wouters (2003) were constructed to answer can therefore also be characterized as being of the how's-that type. Compared to the calibration approach, the purpose is not just to answer a single well-posted question but rather to be a tool that can be applied to various macroeconomic domains.

In addition, we can identify purposes that are different from simply answering questions about phenomena. In the evaluation of DSGE models, a need emerged to formulate objective standards of model performance (Fernández-Villaverde & Guerrón-Quintana, 2021). Comparing the performance of models that seek to answer similar questions is seen as important because different models may provide different answers to the same question. Furthermore, one would like to know whether the alteration of a model structure leads to improved performance to guide potential future research.

In terms of theory, the shift toward the estimation approach is accompanied by a shift from real business-cycle theory (RBC) toward new Keynesian theory (Fernández-Villaverde & Guerrón-Quintana, 2021). New Keynesian theory introduces nominal rigidities (sticky prices in the short run), but at its core it is still comparable in many aspects to the RBC theory. It starts from consumers and firms that maximize their inter-temporal utility. This implies that the representational core of estimated DSGE models is still strictly microeconomic, which seeks to enhance the invariant nature of the model. In these aspects, overlap is found in the theoretical and the technical criteria between the estimation and calibration approaches.

Let us now turn our attention to what is, in my view, a key difference between the calibration and estimation approaches, namely the way in which the model is parametrized. Instead of using calibration to quantify the model's parameters, in the latter, maximum likelihood estimation is applied. Before delving into the relationship between this method of parametrization and the validation criteria, let us first discuss what such an estimation procedure entails. Maximum likelihood estimation as a means to parametrize stochastic macroeconomic models was first demonstrated in Sargent (1989). More recently, Fernández-Villaverde and Guerrón-Quintana (2021) provides a general framework of this methodology. It is helpful to discuss it in more detail to attain a sense of its methodological underpinnings. The framework introduces a state-space representation of the DSGE model and the data, the first equation of which is as follows::

$$S_t = h(S_{t-1}, W_t; \Sigma), \tag{3.1}$$

where S_t is a vector of state variables at time t. In most DSGE models, such state variables will include the level of consumption in time t and the price level at time t. W_t is a vector of stochastic shocks, which are often distributed normally and are independent and identically distributed. Σ is the vector of parameters that needs to be estimated. Since W_t is stochastic, we can rewrite Equation 3.1 as a conditional probability distribution: $p(S_t|S_{t-1}; \Sigma)$. The second equation is:

$$D_t = g(S_t, V_t; \Sigma), \tag{3.2}$$

where D_t is a vector of observables. V_t is a vector of shocks outside the model, such as measurement errors on observables. This equation can be rewritten as $p(D_t|S_t; \Sigma)$. Equation 3.1 can be substituted into Equation 3.2, which yields:

$$D_t = g(h(S_{t-1}, W_t, \Sigma), V_t; \Sigma) \tag{3.3}$$

This can be represented as a probability function of observables, conditional on the vector of state variables at $t-1$: $p(D_t|S_{t-1};\Sigma)$. Note that this probability function is the mDGP. Using this conditional probability function we can take a sequence of observations $d^T = \{y_1, y_2, ...y_T\}$ as input and estimate the probability p of that sequence occurring for a given Σ. The combination of parameter values Σ that maximizes p are the resulting estimated parameters. The functions $h(.)$ and $g(.)$ are generally unknown and cannot be found explicitly. Instead, numeric simulations are used to estimate Σ.

The application of maximum likelihood estimation for the parametrization of DSGE models is a technical validation criterion. The question is: How does this criterion follow from the purpose of the model? According to Fernández-Villaverde and Guerrón-Quintana (2021), three main channels present the three key advantages of the estimation approach over the calibration approach.

First, recall that the purpose of estimated DSGE models is often to answer multiple distinct questions. The model purpose is thus quite broad. This requires a model with a more complex structure compared to calibrated DGSE models, which often have a more limited purpose. A more complex model structure implies more model parameters to quantify – with more degrees of freedom to address. When calibration is applied in this case, the problem of underdetermination may occur (Fernández-Villaverde & Guerrón-Quintana, 2021); multiple combinations of parameters are consistent with the reproduction of the set of relevant stylized facts. The problem with the calibration approach is the lack of data points (in the form of stylized facts) to which the model data can be fitted. Direct calibration (importing parameter values from microeconomic data) can only mediate this issue to an extent, since many parameters have no clear direct measurements (see Section 2.1.1 for an analogous discussion on calibration in the case of MABMs). Estimation does not typically incorporate specific and stable data characteristics but rather all the available data points. This allows for the identification of a unique set of optimal parameter values.

Second, the purpose of an estimated DSGE model may be to provide quantitative predictions of the variables of interest. This purpose requires the model to provide accurate forecasts (under some accepted level of uncertainty) of all future variations of the variables of interest. In turn, this requires the model to be fitted to all observed variations of the variables of interest, which is what maximum likelihood methods are able to do. Calibration models that are constructed only to reproduce a set of stylized facts are generally unable to provide quantitative forecasts of this kind.

The third way in which model purpose relates to the estimation approach is the desire to establish an objective standard for comparing the performance of different models that have a similar primary purpose. In the estimation approach, this comes in the form of the likelihood score. In the calibration approach, comparing model output and empirical data is more casual, and generally no rigid measure of fit exists.

How does the technical criterion of applying the estimation approach affect the model scope? Recall that the output scope is defined as the degree to which overlap exists between the data generated by the rwDGP and the mDGP. The estimation and calibration approaches require models that have different output scopes. Specifically, the calibration approach seeks out stable characteristics of the data generated by the rwDGP. The model is partially parametrized by selecting those parameters that allow the model to reproduce these stable characteristics. This implies that the output scope is limited to the selected stable characteristics. On the contrary, the type of maximum likelihood methods used in the estimation approach typically do not select specific and stable characteristics on the data; instead, the mDGP is described as a conditional probability function of observables, and all data points are assigned a particular probability value. This implies that any observed variation in a particular variable co-determines the resulting estimated parameters. The parameters are thus selected based on the overlap of selected variables between the data generated by the mDGP and all variations in the data generated by the rwDGP. The estimation approach thus requires a target scope that is more encompassing than what is required in the calibration approach.

This leads to an interesting question regarding how we can construct models that yield an mDGP that is able to capture a large degree of the observed variation of a particular set of variables. In my view, there are two ways of accomplishing this, both of which revolve around specific requirements for the structural scope.

The first and more straightforward method is to construct a model that yields an mDGP that embeds a sufficient number of structural elements while maintaining a relatively high degree of evenness in the nature of the relationship model structure in the mDGP and the structure in the rwDGP. This means that the majority of the model elements are to be primarily interpreted in a similar way in relation to the structure of the rwDGP. We will label this as a uniform structural scope. For example, if the model purpose is to provide an answer to a question that requires an explanation of a phenomenon, the model structural elements should be interpreted as, at least to some degree, injective (one-to-one) representations of the real-world structure (see Section 1.2 for a discussion of representation). In this case, an expansion of the output scope thus comes down to increasing the number of representational elements in the model structure. An important example here is the 1950s Cowles Commission approach to macroeconomic modeling. This approach consists of estimating simultaneous equation systems with macroeconomic variables. To establish invariant relationships between variables, the model scope was expanded by incorporating all macroeconomic relationships that were deemed important from a theoretical point of view. For this reason, Cowles Commission-type models could become very large, consisting of hundreds of equations. An example is the Brookings model of the United States (Dusenberry et al., 1965), which comprises over 400 equations. A downside of this approach was that the intelligibility of the model suffered (Boumans, 2009).

Another example are vector autoregressive (VAR) models, which consist of a vector of macroeconomic variables that are regressed on the lags of the same variables. Such a model can, in the context of our framework, be regarded as non-representational (Sims, 1980). Furthermore, the general purpose of VAR models is to provide quantitative predictions. VAR models are generally able to match the observed variation well, without claiming that its representational elements are to be interpreted as representational of the real-world structure.

Accordingly, what the VAR approach and the Cowles Commission approach have in common is that they have a large output scope while presuming a uniform structural scope. In the case of VAR models, the majority of elements are primarily to be interpreted as non-representational. In the case of the Cowles Commission approach, the majority of elements are to be primarily interpreted as representational. Neither models rely on a combination of representational and non-representational elements.

The second way to capture a large degree of the observed variation is to construct a model that yields an mDGP that has a hybrid structural scope. This implies that some parts of the model structure require a different type of interpretation than others, where the model crucially relies on both types to fulfill its purpose. I argue that the way in which the target scope is expanded in estimated DSGE models is through a hybrid structural scope. The structure of estimated DSGE models consists of a representational core and a non-representational periphery. As described above, the structure of the representational core is deduced from economic agents that optimize utility under certain constraints. The representational core often consists of a few equations that describe some of the macroeconomic relationships. These are generally used to describe how the economy moves back to equilibrium after a shock. In line with our discussion of the model purpose of estimated DSGE models, the representational core of the model is used to answer how's-that questions.

However, the representational core lacks the output scope needed to yield satisfactory estimation results when maximum likelihood is applied. That is, the fit is insufficient between the model data and the empirical data. The variation in the model data accounts only for a relatively small proportion of the empirical data. Geweke et al. (1999) provide a discussion of how early attempts to estimate DSGE models failed for this specific reason. The likelihood score is very low if the output scope is insufficient. In response, the innovation introduced in Smets and Wouters (2003) is a model that takes the representational core and supplements it with stochastic terms and lags of those terms that are non-representational. In isolation, the non-representational periphery is not suited to provide economic understanding. Rather, using the language of time-series econometrics, it is a moving average process. Typically, such processes are used to provide quantitative predictions about future values of economics variables (see, for example, Holt (2004)). To connect this to the types of questions we have distinguished above, we can see that moving average models are used to answer how-much questions. The purpose and interpretation of the stochastic periphery is therefore different from that of the representational core. This implies that the model is hybrid in terms of its structural scope.

This hybrid form yields a model with a relatively parsimonious structure compared to, for example, Cowles Commission-type models, while at the same time maintaining a sufficient output scope such that the model can be estimated rather than calibrated. At the same time, because of its representational core, the model's answer to questions maintains an explanatory status. In the next section, however, I will discuss whether the correctness of such answers can be adequately assessed.

3.4 Analyzing the Hybrid Model Critique

Now that we have seen how the calibrated DSGE approach and the estimated DSGE approach can be understood from the perspective of the model construction framework, we can dive deeper to discuss the implications of these approaches and gain a more fundamental understanding of the hybrid model critique.

In my view, the essence of the hybrid model critique can be understood as a criticism of a model structure that disrupts the relationship between the validation test of the model structure and its actual validity. Recall that validation is the assessment of a model's ability to answer a particular question. Disruption of this process means that it is no longer able to assess whether the answer to the question is correct. The critique, as formulated by Chari et al. (2009) and De Grauwe (2012), also starts from this assertion; doubt emerges over the validity of the model, despite the fact that the model has passed empirical validation tests. The question is, therefore, why the relationship between validation tests and the model's validity is disrupted as a result of the hybrid model structure.

First, it is important to emphasize that we have defined a hybrid structure as a hybrid structural scope, meaning that the relationship between the model structure and the real-world structure is not uniform but hybrid. As discussed before in the previous chapter (Section 2.4.1), three types of questions require distinct validation criteria that, in turn, impose different requirements on the structural scope. We have established that estimated DSGE models have a representational core that is suitable for how's-that questions, as well as a non-representational periphery that is suitable only for answering how-much questions. To connect these questions to the types of models in Section 2.4.3, we can say that the representational core is a grey-box model, and the non-representational periphery is a black-box model. As we have discussed, these model types are associated with different types of structure validation.

This non-representational periphery of the model is not subject to indirect structure validation, because the relationship between the model structure is not supposed to be in a representative relationship with the real-world structure. This implies that only the model's ability to reproduce its target is to be assessed through validation (see target validation in Section 2.4.2). For example, if the purpose of the model is to predict validation in the next month, we should test whether it is able to do so without needing to address the structure by which the model functions. Hence, for the non-representational periphery of estimated DSGE models, there is no need to test the overlap between the model and real-world structures.

On the other hand, the representational core of the model structure is subject to structural validation in addition to target validation. This is generally done by comparing a wide range of model outputs to empirical data. If the model is able to provide the right answers to questions to which we know the answers, we can be confident that the model structure behaves at least as-if the real-world structure within a relevant domain. The model structure is thus validated in terms of the output of the model as a whole.

On the other hand, the representational core of the model structure is subject to structure validation in addition to target validation. This is generally done by comparing a wide range of model outputs to empirical data. If the model is able to provide the right answers to questions we know the answers to, we can be confident that the model structure behaves at least as-if the real world structure within a relevant domain. The model structure is thus validated in terms of the output of the model as a whole.

The problem with the validation procedures described in the previous paragraph is that structural validation is only applicable to the representational core of the model but makes use of the model output generated by the model as a whole. The structural elements of the representational core and non-representational periphery interact and work in concert, generating model output that we then compare to the data. The output of the model as a whole cannot therefore be interpreted as being generated from structural elements that can be interpreted as representational. Some of the structural elements of the model that are responsible for generating the model output lack a representational interpretation. This is equivalent to the specific critique in Chari et al. (2009) that, given the lack of a representational interpretation of shocks, the model is not yet suited for policy guidance.

The essence of the issue is that when a model with a hybrid structural scope passes validation tests based on an output scope produced by a partially non-representational structure, we cannot interpret the passing of these tests as a signal that the representational core of the model is valid. In other words, we cannot interpret it as a signal that the model structure behaves at least as-if the structure of the rwDGP.

Note that in the introduction of this chapter, we stated that the notion of representational here should be thought of as different from not relying on artifactual elements. In fact, the structure of the representational core in the case of grey-box models relies to a large extent on what we labeled artifactual elements in Section 1.2. Rather, representational structures explicitly describe mechanisms that have an economic interpretation, whereas a non-representational structure is open to multiple possible economic interpretations. Mechanisms that have an economic interpretation may still rely on artifactual information to a large degree, in that they cannot be

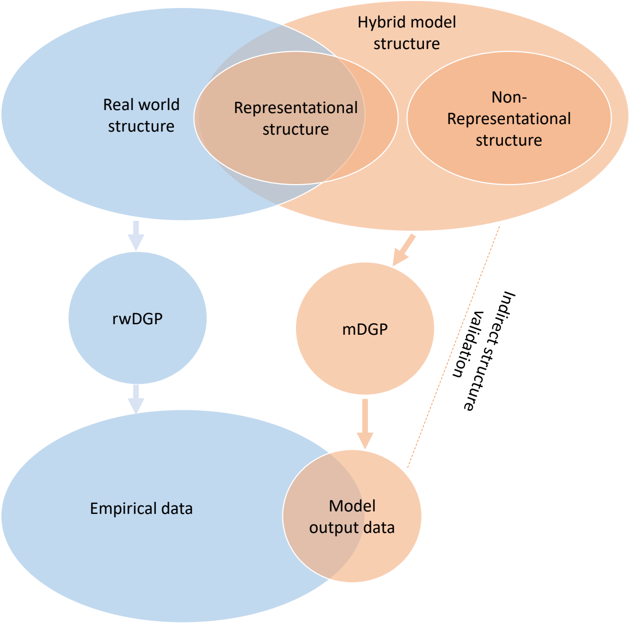

Hybrid model structure
Real world structure
Representational structure
Non-Representational structure
rwDGP
mDGP
Indirect structure validation
Empirical data
Model output data

Figure 3.2: Validation of hybrid model structure

inferred from data or theory alone. The fundamental difference between representational and non-representational structures in relation to validation is whether the structure is of interest and used to provide some sort of economic understanding. Importantly, a structure can still provide understanding even if it relies on artifactual elements to a large degree.

Figure 3.2 presents a schematic overview of what I have just described: Both the representational core and the non-representational periphery are part of the model structure that makes up the mDGP. This mDGP generates a uniform output. In DSGE models, the output scope is used to validate the structural scope.

A defense of the hybrid model structure may be that the representational core and stochastic periphery are, in fact, independent. This is often done by distinguishing between an endogenous structure – that which is within the realm of economics – and an exogenous structure – outside forces that affect the economy but are outside of what can be explained economically. In the hybrid model structure, the representational core is endogenous while the non-representational periphery is exogenous. Perhaps the exogenous structure represents (geo-)political forces, or relevant elements of the natural world. One can then argue that if this is the case, the representational core is not affected structurally by the non-representational elements. If this division between endogenous and exogenous factors is an accurate and complete description of the structure of the economy, it would allow us to match the representational core and stochastic periphery independently of the observed variation in the data. To put it differently, part of the observed variation can be attributed to what we would expect given that the representational core of the model and any remaining variation can be attributed to the stochastic periphery. This would allow for the uniform output scope to be separated in line with the representational scope.

Although this is a strong assumption, it becomes more plausible if the contribution of the stochastic periphery to the total model dynamics is small. In modeling strategies like the Cowles Commission approach, the stochastic periphery is made small by making the representational core as encompassing as possible (Boumans, 2007). We have seen that with calibrated DSGE models, the stochastic periphery is made small by limiting the data used to parametrize the model. In estimated DSGE models, however, the contribution of the stochastic periphery to the observed dynamics is not small. Both Chari et al. (2009) and De Grauwe (2012) argue that in contemporary DSGE models, the exogenous, non-representational elements of the model are responsible to a large degree for the model dynamics. Following the endogenous/exogenous dichotomy, this implies that most of the economic dynamics we observe are outside the realm of economics. Given the importance of the non-representational elements, it is unlikely that the representational elements are fundamentally independent. More likely, the non-representational elements are a stand-in for the relevant economic dynamics that are absent from the relatively simple representational core. Problematically, this includes elements that may be relevant from a policy point of view. A similar point is made in Chari et al. (2009), where the effect of non-representational elements can have multiple representational interpretations with different policy implications.

Based on the insights of the model construction framework, the hybrid model critique is justified. Importantly, the hybrid model critique should be understood as a weakening of the validity signal derived from empirical validation tests. As we have seen, the hybrid model structure was introduced, ultimately, due to a shift and broadening of the model purposes: from a limited purpose of answering a well-posted question about a particular economic phenomenon toward models that can describe a multitude of phenomena and make quantitative predictions, the performance of which can be compared to other existing models. To meet this purpose, the required method of parametrization shifted from calibration toward estimation. In turn, this led to a hybridization of the structural scope to accommodate the estimation methodology. While this has allowed for the construction of more encompassing models, as well as better comparison of model performance, it has come at the cost of validation – that is, a lack of ability to assess the structural correctness of the model.

3.5 Conclusion

In this section, I have introduced the notions of model scope, technical and theoretical validation criteria, and invariance to analyze the shift from the calibration approach to the estimation approach in DSGE models. The calibration approach starts from a limited model purpose. To ensure invariance, it makes use of calibration, which is a technical validation criterion. This allows calibrated DSGE models to ensure invariance while maintaining a limited structural and output scope.

By contrast, the estimation approach to DSGE models starts from a broad purpose, both in the number of questions it should be able to answer and the fact that it should be able to answer multiple types of questions. In addition, its performance is to be compared to that of similar models. The calibration approach is not suitable for these purposes. Instead, it requires a model that is estimated using maximum likelihood methods. Such estimation methods require a model with a sufficiently large output scope. To achieve this, DSGE practitioners have resorted to a hybridization of the model structure. The representational core of estimated DSGE models is similar to the structure of calibrated DSGE models. However, the model is supplemented with a non-representational, stochastic periphery. This has allowed for the construction of models with a broad model purpose while maintaining a relatively simple structure.

The hybridization of the model structure is in some ways an elegant solution to the intended purposes of the estimation approach; yet, in line with the hybrid model critique, it also introduces a lack of ability to assess the structural correctness of the model.

Let me finish by asking what is to be learned from this study. It is important to recognize that inherent tensions persist in terms of the model purposes that may enter the model construction process. As a prime example, purposes that require models with a close fit to the data are at odds , in some aspects, with purposes that require models to be interpreted representationally (Blanchard, 2009). The question that macroeconomics, as a discipline, should pose to itself is whether it should dedicate its resources to constructing models with a broad purpose in an effort to overcome the methodological challenges that come with these models, or should we instead apply a broad set of models, each with a specific and limited purpose.

My view is that there are inherent tension between models that perform well in answering how-much questions and those that aim to answer how's-that or why questions. This is because validation criteria are often at odds with each other. I will provide a more in-depth discussion of this issue in Chapter 6.

Chapter 4

Model Transfer and Universal Patterns: Lessons from the Yule Process

The third and final case in this dissertation is somewhat different from the first two cases. It does not address a specific macroeconomic modeling practice, as in the previous two chapters. Rather, it considers a more general phenomenon in scientific practice: model transfer. Nonetheless, the main interpretive framework of this chapter still centers on validation; as such, it builds on the analysis in the previous chapters.

4.1 Introduction

An observation on the use of models in science is that particular models are used across multiple distinct scientific domains. The term model here refers to the structure of models, which in the case of mathematical models is a mathematical structure. This structure should be understood as abstract, meaning that it does not have any empirical content by itself.

The observation of a model that is imported into a new domain can be labeled as inter-domain model transfer. For example, the growth process of firms is modeled using the same mathematical structure as the Yule process, which is a model originally developed in evolutionary biology (Simon, 1955).

Such observations contrast with a view of science in which various scientific domains operate in isolation, each using a domain specific methodology. Instead, the observation that particular models are used across multiple distinct scientific domains points to a view of science that is organized through a particular set of methods Humphreys (2004). That is, various distinct scientific disciplines make use of overlapping methods. This does not answer, however, why we would observe such an organizational structure. In fact, it is puzzling when we consider that models are, generally speaking, constructed for a domain-specific purpose: answering a question (Boumans, 2006). Such questions often concern phenomena. For example, how do firms grow in size over time (Simon & Bonini, 1958)? Questions about phenomena are inherently domain specific; they ask about a growth process of, in this example, a specific economic entity, firms. The ability of a model to answer this question is usually built into the model (Boumans, 1999), by shaping the model in such a way that it fulfills relevant validation criteria. Perhaps one would expect that a model shaped by validation criteria that are deemed relevant for a domain specific purpose would always produce a domain specific model, but for some particular models this is not the case.

Such observations contrast with a view of science in which various scientific domains operate in isolation, each using a domain-specific methodology. Instead, the observation that particular models are used across multiple distinct scientific domains points to a view of science that is organized through a particular set of methods Humphreys (2004). That is, various distinct scientific disciplines make use of overlapping methods. However, this does not answer why we would observe such an organizational structure. In fact, it is puzzling when we consider that models are, generally speaking, constructed for a domain-specific purpose: answering a question (Boumans, 2006).

Such questions often concern phenomena. For example, how do firms grow in size over time (Simon & Bonini, 1958)? Questions about phenomena are inherently domain-specific; in this example, they ask about the growth process of a specific economic entity (firms). The ability of a model to answer this question is usually built into the model (Boumans, 1999) by shaping the model in such a way that it fulfills relevant validation criteria. Perhaps one would expect that a model shaped by validation criteria deemed relevant for a domain-specific purpose would always produce a domain-specific model, but this is not the case for some particular models.

The main question that this chapter will seek to answer is: What explains the inter-domain transfer of some models? The observation that some models are transferred across multiple domains implies that these models are somehow considered useful across the domains to which they are applied. Put differently, then, what makes a model useful in the domain for which it was constructed, as well as the domain to which it is transferred? To answer this question, this chapter will introduce a novel framework to understand model transfer. The framework builds on the notion that a model is constructed to fulfill various types of validation criteria, as discussed in Sections 2.4.2 and 3.2. Given this framework, I will show that inter-domain model transfer can be explained as overlap between validation criteria across domains. In particular, special attention will be paid to the overlap between phenomenological validation criteria.

To explain how this overlap can occur, I will introduce the notion of universal patterns. Universal patterns are abstract structures that, when coupled with empirical content, can be applied to multiple distinct domains. Empirical content refers to the information that relates an abstract structure to objects that can be observed empirically (Humphreys, 2019). To illustrate my analysis, I will discuss a case study of model transfer. The study concerns the Yule process, a model first developed in evolutionary biology (Yule, 1925) and later transferred to various other systems, including the growth of firms (Simon, 1955).

In the existing literature, we can distinguish three main accounts that seek to explain model transfer (Knuuttila & Loettgers, 2020): analogies (Hesse, 1966), which attribute model transfer to similarity relationships between phenomena, formal templates (Humphreys, 2019), which attribute model transfer mainly to overlap in construction assumptions, and model templates (Knuuttila & Loettgers, 2016), which attribute model transfer to overlap in conceptual features. Each of these accounts embed a notion of inter-domain model usefulness. They point to particular aspects of models that allow scientists to reuse these models across distinct domains. Although valuable, however, I will argue that these accounts do not give a complete enough description of what it is that makes a model considered useful in practice.

Examining models as analogies is discussed in Hesse (1966), among others. In this account, models derive utility from their similarity relations with the phenomenon of interest. Hesse (1966) distinguishes between positive, negative, and neutral analogies. In the context of models, positive analogies represent the aspects of the phenomenon of interest and the model's features that overlap; negative analogies are those that do not overlap. Neutral analogies are the aspects for which this overlap is yet to be determined and are thus what makes the model potentially useful to learn about the phenomenon of interest. For a model to be useful, it must therefore be a positive analogy of the phenomenon of interest in that particular domain to some degree. In the case of model transfer, this implies that the features of the model provide a positive analogy in both the original and new domains. This is likely to be the case when a similarity relation exists between the targeted phenomena of the different domains. If we consider a model of genera growth in biological evolution that is also used as a model for firm growth (as in Simon and Bonini (1958)), it is likely that certain features serve as analogies to genera growth in biological evolution as well as firm growth. Importantly, such features cannot be domain-specific and are thus to some extent abstract. As we will see in this chapter's case study, one of these features is proportional growth, which can serve as an analogy for how both genera and firms grow. What is transferred, according to this account, is thus an analogy that applies to multiple domains. However, this still leaves open why it is that certain abstract features can serve as positive analogies in multiple domains. Furthermore, as noted in Humphreys (2019), such analogies can often be made to fit in a domain opportunistically. As a result, simply looking at model transfer in the context of analogies may not always yield a satisfactory account of model transfer.

A different view comes from Humphreys (2004), who posits the idea of a computational template. A computational template is a computational structure that can be adjusted for use as a model in distinct domains. The utility of using this computational template, and the explanation as to why some models become templates, are favorable analytical-tractability properties. The template should also be flexible; it should be open to adjustments so that it can be made to fit various distinct domains. This view of model transfer, however, was originally forwarded to be applicable to computational models.

More recently, (Humphreys, 2019) has provided an extension of this account. This view characterizes what is being transferred as a formal template. In this account, the usefulness of a model is essentially determined by the correctness of a model's construction assumptions. Model transfer here is therefore enabled by the correctness of the construction assumptions in the original and new domains on a more abstract formal level. If a construction assumption is a linear relationship between two variables, this assumption should hold in both domains; what is transferred in essence is thus not an analogy but a "correct" formal structure with favorable formal properties. However, Knuuttila and Loettgers (2020) states that solely considering formal properties is not a complete explanation because it does not explain why some models are transferred between domains widely and others are not. Many models that are successfully used within a particular domain will have favorable formal properties, such as analytical tractability, yet only a few are transferred across domains.

Another important addition to the model transfer literature is (Knuuttila & Loettgers, 2016), in which the concept of a model template is introduced as a template with favorable formal properties coupled with general conceptual features. These conceptual features suggest how to theorize about the phenomenon described by the model. This implies that model transfer is enabled when the conceptual features embedded in the template are deemed useful tools for theorizing in both the original and new domains. Examples of such conceptual features are given in Knuuttila and Loettgers (2020), including phase transitions and local interactions. The account of model templates points to a particular source of model usefulness that allows us to explain some instances of model transfer. In my view, however, the account is most applicable to the methods and conceptual notions present in complexity science and is therefore limited in its scope of application.

The essential difference between the account of model transfer put forward in this chapter is that it does not rely on a particular epistemological account of model usefulness. Instead, rather than explaining what makes a model useful, I will take a more empirical approach and explore what makes a model *considered* to be useful in observed scientific practice. In my view, this approach results in an account of model transfer that is a closer match to scientific practice and which therefore covers a wider range of model-transfer cases. It also does not rely on a particular epistemological view of model usefulness. Furthermore, it highlights an enabling factor of model transfer that is not explicitly present in the accounts of model transfer discussed above, namely universal patterns. The account presented here is also general in the sense that it subsumes the existing accounts of model transfer to some extent.

To specify the aforementioned criteria of model usefulness, I build on the concept of model validation and the various validation criteria discussed extensively in the previous chapters (Sections 2.4 and 3.2). From the point of view of validation, model transfer is enabled by satisfactory validation in the original and the new domains, which is in turn enabled by overlapping validation criteria. In this chapter, I argue that empirical validation may play a key role in the transfer process, meaning the assessment of whether the model is able to reproduce relevant facts about phenomena. In such cases, the model that is transferred must be able to reproduce facts about phenomena in both the original and the new domains. Empirical validation as a mechanism of model transfer is supported by the notion of universal patterns, which help us understand why certain models are transferred so widely.

An account of model transfer that also starts from scientific practice can be found in Donhauser (2020). It contrasts two opposing viewpoints regarding the ability of scientists within a particular domain to import knowledge from other scientific domains: incommensurability and voluntarism. Incommensurability is a concept taken from the philosopher of science Thomas Khun in his book *The Structure of Scientific Revolutions* (Kuhn, 1970). It entails that knowledge is domain-specific to such a large degree that knowledge transfer between domains is impossible. At the other end of the spectrum, the notion of voluntarism states that scientists can *choose* a particular epistemological stance as long as certain general conditions are met. Donhauser (2020) argues that incommensurability is not able to explain model transfer, while voluntarism can. As we will see, the idea put forward in this chapter fits neither of these epistemological viewpoints perfectly. Instead, I argue that models are likely to be transferred in the case of overlap in the criteria used to assess model usefulness. The criteria that scientists use do not necessarily have to be the result of voluntary decisions under general conditions; they may also be a function of particular paradigms. As argued in Humphreys (2004), a paradigmatic organization of science is not necessarily domain-specific. Rather, certain methodological strategies span multiple distinct domains.

The reader may associate the notion of model validity with that of robustness as forwarded in Lloyd (2015). Model robustness refers to a degree of insensitivity in a model's ability to reproduce facts about phenomena and to changes in various model assumptions and/or parameter values. Inter-domain model transfer could be seen as robustness with respect to changes in the empirical content of a model. If we change the empirical content of a model (i.e., transfer a model to a new domain), the model is still able to reproduce relevant facts about phenomena. Generally speaking, however, robustness refers to a property of models that reproduce facts about phenomena with the same empirical content. Therefore, to avoid confusion, I will not engage explicitly with the notion of model robustness in relation to model transfer. However, the assessment of model robustness, as it is generally understood, may be subsumed into the more general empirical validation process when relevant. Often, the assessment of model robustness may come in the form of sensitivity analysis – altering parameter values and/or model assumptions and assessing how this affects model output.

4.2 Validation Criteria and Model Transfer

Central to what I have argued in the previous chapters (Sections 2.4 and 3.2) is that satisfactory model construction requires the fulfillment of certain validation criteria, and that the model is therefore shaped by its validation criteria. This implies that the model can only be reused in a new domain when it can be validated within this new domain. This is the case if and only if there is overlap in the validation criteria in both the original and the new domains.

Models are the result of a process of construction. They are not just discovered, and they are not a trivial extension of theory. However, the question is whether this construction process is independent of the above-described validation process. In a more traditional view, these processes are considered independent, which roughly means that the validation process starts after the model is constructed. In philosophy of science, this is better known as the distinction between the context of discovery and justification (Boumans, 1999). As shown via case studies in Boumans (1999), the problem with this traditional view is that it is not in line with actual scientific practice. Given that the validation criteria are given by the question that the model is constructed to answer, they are known during the construction process and play an important role in the construction process. Models are constructed in such a way that the model meets these criteria. When the model does not meet these criteria, a "back and forth" process starts in which the model is tweaked and altered until the criteria are met to a sufficient degree. The ability of the model to meet its validation criteria is thus built into the model. The case studied in Boumans (1999), for example, concerns how (in addition to theoretical and mathematical criteria) a micro-founded business cycle model is constructed to reproduce the Phillips curve (the negative relationship between inflation and unemployment), which is a phenomenological criterion.

The account of model construction through the fulfillment of validation criteria applies to models that are constructed from the ground up, as well as models that reuse existing models. Models constructed by recycling existing models are also subject to the various types of validation criteria already discussed.

For models to be acceptable in both the original and new domains, therefore, there must be overlap in the validation criteria. In the framework presented here, overlap in validation criteria is what enables model transfer across distinct domains. The three main types of validation criteria distinguished above are theoretical criteria, mathematical and statistical (technical) criteria (Section 3.2), and phenomenological criteria (Section 2.4.2). For the purposes of this chapter, I will limit the analysis of technical criteria to mathematical criteria. Recall that theoretical criteria require that the model is in line with existing theoretical concepts, mathematical criteria require favorable mathematical properties, and phenomenological criteria require the model to be in line with relevant facts about phenomena.

In the case of theoretical criteria, overlap may be apparent if the core idea of the theory is sufficiently abstract. We can think of certain concepts from evolutionary theory that are considered useful in biology but also in certain sub-fields of economics (Dosi & Nelson, 1994). In the case of mathematical criteria, it is not hard to see that, for example, analytical-tractability criteria may apply across distinct domains. Finally, in the case of overlap in phenomenological criteria, we can think of requiring models to reproduce the same type of empirically cyclical behavior in the original and new domains. While the account of a model template in Knuuttila and Loettgers (2016) can be seen as a vehicle for the fulfillment of theoretical and mathematical criteria, it does not explicitly account for the overlap between phenomenological criteria that enables model transfer. At this point, one may wonder how it is that certain facts about phenomena will remain the same across distinct domains. In the next section, I provide an explanation for the occurrence of overlap in phenomenological criteria.

We may posit that fulfilling these validation criteria shows some similarity relationship between the model and the real-world system; in the case of model transfer, this is evidence of a similarity relation between the targeted real-world system of the original and the new model, which is also implied by an account that examines models as analogies, as in Hesse (1966). However, this depends on the relationship between the fulfillment of validation criteria and the representational value of the model. I argue that it is not useful to consider this relationship for the purpose of this chapter. First, this relationship is complex and uncertain and depends to a large extent on whether one holds a realist or a more instrumentalist stance toward scientific models (Gatti et al., 2018). Second, this relationship depends on the purpose of the model, as discussed in Section 2.4. For black-box models, for example, the sole purpose of the model is to give correct predictions, which implies that the representational value of the model mechanisms is not a relevant criterion of assessment. Not directly engaging with the relationship between validation criteria and the representational value of the model is thus more epistemologically neutral and covers a wider range of model types.

4.3 Universal Patterns

I have stated that overlap in phenomenological criteria should be taken into account to come to a more complete account of model transfer. The question that remains to be answered is: When is this the case? Empirical validation tests generally consist of assessing whether the model is able to reproduce relevant facts about phenomena. Any overlap of phenomenological criteria therefore implies that there is somehow overlap in features of these facts about phenomena. This may seem unlikely, given that facts are by definition tied to what is empirically observed. The distribution of firm size is about a specific domain: firms. Abstract features of such facts, however, may very well appear across multiple distinct domains. These features are what I will label as universal patterns. As we will see, the distribution of firm size follows a particular power law, the Yule distribution, which is a feature of many observed distributions in distinct domains (Simon, 1955).

Let me first elaborate on what I mean exactly by a universal pattern. A pattern can be thought of as an abstract structure. It is abstract because, by itself, the pattern does not have any empirical content, meaning that it is neither empirically true or false (Humphreys, 2019). It is a structure because we perceive it as something structured, as opposed to being unstructured. Typical structures would be geometric shapes, like circles, curves, cycles, and spirals, or, more generally, structures are described as particular mathematical forms. As an example of an abstract structure, we can think of patterns used in knitting; even though the patterns themselves do not refer to anything empirical, we still recognize them as having a specific structure. Patterns can refer to specific facts about phenomena by tying them to specific empirical content. Empirical content, in this sense, refers to the information that relates the abstract structure to the empirically observable facts about phenomena. When the Yule distribution is used as the distribution of genera size, for example, it is coupled with information that gives particular meaning to the shape. A point on the line that is higher than another point on the line means that it represents a genus that is larger in terms of species. Note the four relevant concepts within this description: the pattern, the empirical content, the facts about the phenomenon, and the phenomenon itself.

A pattern is a universal pattern if and only if it can be made to refer to facts about phenomena in multiple domains by changing only the empirical content to which the pattern is coupled. In Figure 4.1, we can see a schematic overview clarifying the relationships between concepts. A single universal pattern can be made to apply

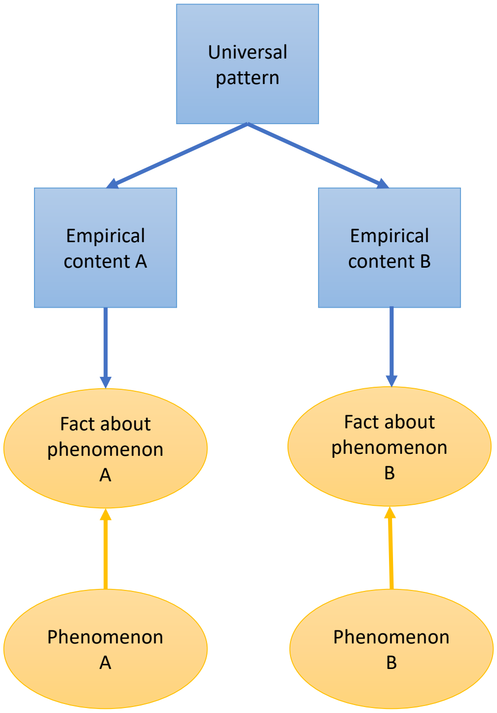


Figure 4.1: Universal patterns and facts about phenomena

both to facts about phenomena A and B by tying it to empirical content A and B, respectively.

The notion of universal patterns put forward here is induced from the observation that certain patterns are observed and used in scientific practice in varying domains. Most straightforwardly, we can think of the Gaussian or normal distribution, which is observed across widely varying domains, such as human height or the weight of loaves of bread (Lyon, 2014). Another example are certain power distributions, such as Zipf's law (Corominas-Murtra & Solé, 2010) or the Yule distribution (Simon, 1955), which are observed in the distribution of city size and the distribution of words in a piece of literature. Universal patterns are not limited to distributions, however; for example, we can think of particular oscillation patterns that are observed in (among many other domains) ecology and economics (Gandolfo, 2008).

Let us now relate the notion of universal patterns more explicitly to phenomenological validation criteria. In order for a model to be transferred across domains, it must be considered useful by the practitioners in both the original and the new domains. This usefulness is considered by assessing whether the model is able to meet certain validation criteria. These validation criteria are built into the model, meaning that the model is shaped by the criteria. For a model to be useful in a domain that is different from the one for which it was originally constructed, the validation criteria should overlap. When phenomenological criteria have played an important role in shaping the original model, it is these criteria that should overlap in the new domain in order for the model to be transferred. This is the case when the phenomenological criteria represent universal patterns. More concretely, when this is the case, models in multiple distinct domains are able to reproduce empirically observed patterns in these domains.

Note that the use of the term pattern here means something different than in Humphreys (2019), where it is used to describe the abstract structure of the model itself. Here, pattern refers to something that has the potential to be observed empirically and to be reproduced by the model. Of course, this means that the pattern is embedded in the model. By working with the model and considering the implications of its assumptions, we can uncover this pattern. In practice, this often involves running numerical simulations of the model. This is in line with the definition of phenomenological output criteria in Section 2.4.2.

The general view is thus that in most modeling exercises, it is desirable to latch the model onto the empirically observable world in some way. The observations we make, and the facts about phenomena that we distill from them, are sometimes structured in specific ways. In such cases, models that are constructed to latch onto phenomena are likely to have a structure that is specific to that observed phenomenon. Such a fact about a phenomenon does not represent a universal pattern. In other instances, however, the facts about phenomena that we distill from our observations are structured in general ways. That is, they embed a pattern that can be made to refer to facts about distinct phenomena – a universal pattern. We are thus confronted with a world in which we observe both specificity and generality. Where we observe specific patterns, there are likely methodological borders. Where we observe universal patterns, there are likely transfers of mathematical forms. This view contributes to an explanation for the observation that some particular models are transferred but not others.

The notion of universal patterns that I have presented here is related to but different from the existing concept of universality. The field that has discussed the notion of universality most explicitly is that of statistical mechanics. In statistical mechanics, universality concerns similarities in the behaviors of diverse systems (Batterman, 2000). Another way in which this is sometimes formulated is that the system-level behavior is independent of elements of the microscopic structure system (Batterman, 2000) . If this is the case, it implies that systems constituted of different objects may still show similar behavior. An example often used is when a magnet is heated to a certain critical temperature, it will lose its magnetism (phase transition). The path between these two states as a function of temperature (coexistence curve) is described by a power function with a critical exponent close to 1/3 (Batterman, 2000). The same functional form and critical exponent is also observed in phase transitions between the fluid and vapor states of matter, including that of water. Clearly, the microscopic structures of water and magnets are different; nonetheless, some properties at the system level are strikingly similar.

The same notion of universality has also been applied to systems outside of chemistry and physics, such as agent-based systems (Parunak, Brueckner, & Savit, 2004) and biological systems (Batterman & Rice, 2014). The power function with a critical exponent close to 1/3 falls within the account of a universal pattern presented here. State transitions in matter and transitions in magnetism are facts about phenomena with distinct empirical content, but they nonetheless express a similar pattern. The account of universal patterns that I have presented, however, does not make any statements about the relationship between the observed pattern and the system by which it is generated. In the notion of statistical mechanisms, universality is a property of a system, the behavior of which comes in the form of widely observed patterns. However, this presupposes that what is observed is strictly tied to the generating system. As I will discuss below, this limits the ways in which we can explain why we observe universal patterns, albeit in a way that is not necessary within the context of model transfer.

Why we observe universal patterns is a fundamental question that requires a full investigation on its own and is thus beyond the scope of the main question of this chapter. Generally, however, we can distinguish between two types of explanations. One comes from the same statistical mechanics notion discussed above, as discussed in Batterman and Rice (2014). It states that systems, although distinct in certain ways, still share abstract fundamental features, such as locality, conservation, and symmetry. Systems that are different in some aspects but share these fundamental features have the same properties – in the form of universal patterns.

This explanation is related to the notion of a causal core, as discussed in (Lloyd, 2015). The causal core consists of those features that are responsible for generating particular output and robust against changes outside this causal core.

For physical systems, this explanation may seem credible, as stated before; however, universal patterns are also observed in diverse social phenomena (Simon, 1955). According to some, such patterns are also the result of abstract fundamental features in the systems by which they are generated. Mandelbrot and Hudson (2007), for example, applies the theory of fractals (Mandelbrot, 1982) as an explanation for the distribution of price changes on stock markets. Fractals are seen by some as a fundamental self-organizing principle of nature (Kurakin, 2011). Somehow, the code of nature is such that distinct systems (even social ones) self-organize into similarly structured patterns.

As an alternative explanation for universal patterns, we can take a more cognitive perspective and question the objective nature of the patterns we observe. As stated before, patterns are abstract structures. What we consider to be structured and unstructured may be shaped by our psychology and limited by our inability to grasp the complexity of the world. This is in line with notions from Gestalt theory, such as those presented in Palmer (1999). Human psychology has a tendency to structure pieces of information into larger information structures in certain ways. The notion of universal patterns that I forward here can be interpreted as ontologically neutral. Here, we are simply addressing the observation that universal patterns are observed by scientists, which therefore partially determines which models we consider useful.

4.4 The Yule Process: A Case Study

To illustrate the account described above, I would like to discuss the Yule process and the universal pattern that can be derived from it: the Yule distribution. I have chosen this example of model transfer because an explicit account exists of how this model was constructed in Yule (1925) for its original context, as well as how the model was later used as a basis for the construction of models in other domains (Simon, 1955). More recently, the Yule process has formed the basis for many models that concern preferential attachment (Abbasi, Hossain, & Leydesdorff, 2012), which is a central notion in network theory (Newman, 2001).

4.4.1 Yule Process: Evolutionary Origins

George Undy Yule (1871–1951) is known as a pioneer in the field of statistics. The model that is the subject of this case study is called the Yule process. The distribution that can be derived from this process has been labeled the Yule distribution, which is perhaps his most well-known scientific contribution (Edwards, 2001). A short history of the development of the model can be found in Bacaër (2011), on which the analysis below is partially based.

Yule developed his model in response to observations made by the botanist J. C. Willis (1868–1958) in evolutionary biology. The subject concerns the distributions observed in taxonomy, a biological classification scheme with a hierarchical structure in which organisms are grouped together based on common characteristics. The system is hierarchical in the sense that classifications with a higher taxonomic rank are more general, thus embedding a classification of more specific lower taxonomic ranks. The observations made by Willis span two such ranks: specie and the more general rank of genus. A given genus thus contains multiple species that have certain features in common at the genus level but differ at the species level. The suborder of -Snakes-, for example, contains many more specific genera, such as -Boa-, which in turn contains the specie of -Boa Constrictor-.

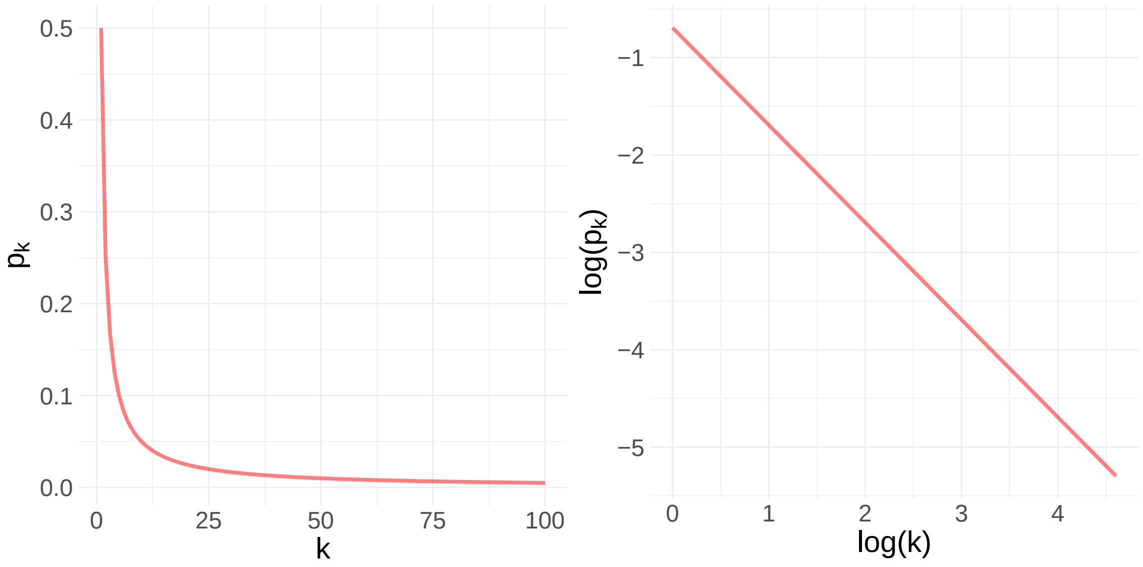
For several different organisms (animals and plants), Willis collected data on the number of genera that contain a given number of species. In this context, we can say that the size of a genus is determined by the number of species it contains. By tabulating this data, an interesting distribution emerged; many genera contained one specie (size one), some genera were larger, and a few genera were very large and contained more than 100 species (size 100). What was also striking is that this pattern appeared to emerge in both animals and plants.

Yule, who trained as a statistician under Karl Pearson, suggested plotting the data on a log-log scale. This revealed that the logarithm of the fraction of genera containing k species, $log(p_k)$, decreased approximately linearly with $log(k)$. This implies that$\alpha > 0$ and $\beta > 0$, such that the probability density function of genera size can be written as:

$$p_k \propto \alpha k^{-\beta} \tag{4.1}$$

Which can be rewritten as:

$$\log p_k \propto \log(\alpha) - \beta \log k \tag{4.2}$$



(a) $p_k = 0.5k^{-1}$

(b) $\log p_k = \log(0.5) - \log k$

Figure 4.2: Power Law for $\alpha = 0.5$ and $\beta = 1$

Figure 4.2 plots both equations for arbitrary parameters. In addition, Willis made observations regarding the age of a genus and its size and found that larger genera were older on average, evolutionarily speaking.

Yule was interested in providing a mathematical model based on evolutionary theory that was able to reproduce Equation 4.1 and, in addition, to explain the observation made by Willis that the larger genera were also older. In Yule (1925), he provided this model, stating its purpose as follows:

The Further question arises, what is the frequency distribution, as the statistician terms it, of the sizes of these N genera which all started as monotonic genera from primordial species at zero time, after any given time has elapsed?(Yule, 1925)

This purpose encapsulated the desire to generate the distribution of genera size while linking genera size to evolutionary age. From the outset, there were thus some clear validation criteria that are in line with the ones I have discussed previously: a theoretical criterion, in that the model assumptions must roughly agree with evolutionary theory, and a more explicitly phenomenological criterion, in that the model must able to reproduce a distribution that is linear on a log-log scale.

Let us now take a look at how Yule managed to construct a model that reproduces a frequency distribution that is in agreement with these "known facts." The two fundamental entities in this model are species and the genera to which they belong. Consider how these two entities grow over time. The total number of genera is labeled n. Each genus has a size k that is determined by the number of species belonging to each genus at a point in time. At each time step, m species in total are added to the existing genera. After these m species have been added, a new genus is added to the existing genera; this new genus starts with $k = 1$ Subsequently, the total number of species has increased by $m + 1$ (m plus the specie associated with the new genus); $m + 1$ new species appear for each new genus added, implying that the average number of species per genus is $m+1$. With each time step, n is increased by 1. This implies that the number of time steps can be represented by the total number of genera n. $p_{k,n}$ is the fraction of genera with k species when the total number of genera is n. The total number of genera with k at n is $np_{k,n}$.

At this point, the probability of a species being added to an existing genus becomes crucial. This probability is taken to be proportional to the size of the genus, such that if we have a genus with k_i species, the probability of a specie being added to this genus is given by the number of species belonging to genus i over the total number of species:

$$\frac{k_i}{n(m+1)}. \tag{4.3}$$

We now have all the ingredients of the model. In short, the model consists of two main elements: constant genera growth and proportional specie growth. The question to ask now is: Where do these ingredients come from?

Part of the response is a general knowledge of evolutionary theory. In the introduction to his chapter, Yule discusses two opposing views regarding how evolution occurs that were relevant at the time. The first is what Yule labels the "Darwinian view," which assumes that differences in species and genera arise through cumulative small mutations (continuous variation) and that species necessarily die out. The "mutational view" assumes that large mutations may occur "at once *per saltum*," as Yule phrases it, meaning with large jumps (discontinuous variation).

It may seem that the type of mutation described in the model, as well as the assumption that species do not die out, is more in line with mutationalism. However, Yule is well-known for his opposition to mutationalism, most prominently in Yule (1902).

To ensure that his assumptions do not disagree with the Darwinian view, Yule provides an explanation of how the model's assumptions should be interpreted. First, mutations in his model are limited to "viable mutations," such that the model does not formally contradict the dying-out of species. Second, Yule notes that given a long enough time horizon, small continuous mutations accumulate into changes that may appear discontinuous. The time horizon in the model should thus be interpreted as long enough for such small mutations to accumulate into something that would be classified as a new specie or a new genus. Clearly, then, an effort was made to position the model within the context of existing evolutionary theory. Such considerations provide us with an example of how the ability to meet theoretical criteria is built into a given model.

However, the model proposed by Yule was certainly not an injective (one-to-one) mapping of evolutionary theory. Interestingly, behind proportional growth is the assumption that the probability of creating a new specie is the same for each individual species, regardless of genus and time. This implies that larger genera will grow at a higher rate in absolute terms. Regarding this assumption, Yule writes as follows:

The assumption that the chances of specific (or generic) mutation are identical for all forms within the group considered are constant for all time are unlikely to be in accordance with the facts, but have to be made to simplify the work. (Yule, 1925)

Why did Yule make this non-factual assumption? Here, we enter the realm of analytical-tractability/mathematical criteria: Introducing heterogeneity in the rates at which hundreds of species and genera evolve would undoubtedly complicate the model's computational structure and may hamper the degree to which the model would enhance understanding. In addition, it could be that such a model can only be implemented through computer simulation, which was not a tool available to Yule. To convince the reader of the correctness of this assumption, Yule points not to evolutionary theory but to empirical facts that the model must be able to reproduce – the phenomenological criteria:

In so far as the deductions do not agree with known facts the assumptions are probably incorrect or incomplete. In so far as we find agreement, or the more nearly we find the agreement, the assumptions are probably correct.(Yule, 1925)

Indeed, the model proposed by Yule is able to reproduce the frequency distribution of genera:

So for as the graphic test goes, accordingly, the theory gives very well indeed precisely the form of the distribution required.(Yule, 1925)

From the outset, before any formal derivation, we can see that the constant addition of small genera, coupled with a proportional growth of species, would generate a distribution with some very large genera and many smaller ones; to put it mathematically, a skewed distribution. Starting with only genera with $k = 1$, some genera will by chance grow slightly larger than others. These larger genera will then have a higher probability of growing even larger (following Equation 4.3), and so on.

The above description of the construction of the Yule process shows how the model is shaped by a balancing act between three validation criteria. Specifically, the model had to be (to some extent) in line with notions from evolutionary theory, it had to be solvable analytically, and it needed to reproduce the observed statistical distribution. It was these criteria that served as Yule's standards for model usefulness. This shows that the Yule process is a model constructed for a specific domain and shaped by the validation criteria within this domain.

4.4.2 The Yule Process as a Model for Firm Growth

How was the structure of the Yule process used as a basis for the construction of models in other domains? In the above analysis, we have established that overlap in validation criteria between domains is necessary for models to be useful in multiple domains. Let us look, therefore, at which considerations were most important in the selection of the Yule process as a basis for constructing models in a new domain.

The Yule process has been used to model the processes of many different subjects (Simon, 1955). As an example, we will look at how it was first applied to model the distribution of firm size in Simon and Bonini (1958). Let me first provide some background on the scientific discussions regarding models of firm size in this period. At that time, it had long been observed that the distribution of firm size was heavily skewed (Gibrat, 1931), implying a distribution of some very large firms and many smaller firms. The non-normality of this distribution was seen as evidence of the non-trivial nature of the growth process. This observation brought with it a dissatisfaction with standard economic theory because it was unable to make predictions regarding the distribution of firm size (Simon & Bonini, 1958). Born from this dissatisfaction, the goal of Simon and Bonini (1958) was to provide a model that could generate the observed distribution of firm size. From the start, therefore, the model construction was aimed at a phenomenological criterion.

Simon and Bonini (1958) starts with the assertion that to generate the distribution of the type observed in firm size, the law of proportional effect is an essential ingredient for the model. The law of proportional effect was first introduced by Gibrat (1931) and entails that growth is proportional to size. It is the same structure labeled by Yule as proportional growth. In the case of firms, this would mean that the same percentage of growth rates applies to firms of different sizes. This implies that larger firms grow faster in absolute terms. Concretely, this means that the expected percentage return on investment is not a function of firm size. Computationally, this is in line with growth in the original Yule process, in which larger genera will also grow at higher absolute rates. However, this was not enough to narrow down the appropriate model to one. Simon and Bonini (1958) states that there may be multiple distinct growth processes (models) that will generate the type of distribution skewness observed empirically, as long as proportional growth is incorporated:

If we incorporate the law of proportionate effect in the transition matrix of a stochastic process, then, for any reasonable range of assumptions, the resulting steady-state distribution of the process will be a highly skewed distribution, much like the skewed distribution of that have been so often observed for economic variates. In fact, by introducing some simple variations into the assumptions of the stochastic model - but retaining the law of proportionate effect as a central feature of it - we can generate the log-normal distribution, the Pareto distribution, the Yule distribution, Fisher's log distribution and others - all bearing a family resemblance through their skewness.

Proportional growth was thus deemed essential for generating the type of distribution observed for the size of firms. However, this still left open a range of skewed distributions and processes that generate them. To narrow down the growth process further, Simon and Bonini (1958) examined more closely the characteristics of the observed distribution of firm size:

The log-normal function has most often been fitted to the data and generally fits quite well. It has usually been noticed, however, that the observed frequencies exceed the theoretical in the upper tail and that the Pareto distribution fits better than the log-normal in that region. The observation suggests that the stochastic mechanisms proposed in the previous section are the appropriate ones and that the data should be fitted with the Yule Distribution. (Simon & Bonini, 1958)

The observed pattern is thus one of a particular shape: it is log-normal except for the upper tail, which is Pareto distributed. These two characteristics are consistent with the pattern of the Yule distribution. To reproduce this pattern, Simon and Bonini (1958) incorporates the second essential ingredient of the Yule process: the constant entry of new small firms. In this way, Simon and Bonini (1958) arrives at a model with the same structure as the original model that is able to meet the validation criteria within the new domain.

4.4.3 Overlapping Validation Criteria

Where can we find overlap in the validation criteria between the original and new domain? First, if we look at theoretical criteria, we do not see strong indications of overlap. The evolutionary theory that served as a criterion in the original construction of the Yule Process did not play an explicit role when the model was applied to firms. In Simon and Bonini (1958) we see that theoretical criteria did not seem to play a big role altogether. Rather, Simon and Bonini (1958) is partially born out of a dissatisfaction with the inability of microeconomic theory to explain certain empirical patterns.

Second, both models contained an at least implicit mathematical criterion of analytical tractability. The Yule process was a good candidate because the model was shown in Yule (1925) to fulfill this criterion. In line with Knuuttila and Loettgers (2020), this criterion is fulfilled by countless models and is not enough to narrow things down to a particular model. By itself, then, it is not a complete explanation of why the Yule process was transferred to the new domain.

Third is the overlap between the pattern observed in the distribution of genera size and the pattern observed in the distribution of firm size. It was this pattern – a certain shape – that enabled the model of the Yule process to be considered useful in both domains.

4.5 Conclusion

What explains inter-domain model transfer in science? In this chapter, I have put forward an account of model transfer that begins with the construction process of models in practice. In practice, models are constructed such that they meet relevant validation criteria. These criteria can be theoretical, mathematical, or phenomenological in nature. The models are shaped by these criteria. In this sense, a model can thus be seen as a containing structure that meets certain criteria.

If such criteria are domain-specific, the model will only transfer within the original domain of construction. If, however, the validation criteria also apply to other domains to a large enough extent, the model may also be considered a useful tool in these domains. Inter-domain overlap in theoretical criteria applies in cases where the core of the theory in question is sufficiently abstract, such as complexity science. Mathematical criteria play an important role in shaping many models, and these criteria will often overlap between domains – analytical tractability, for example. I agree with Knuuttila and Loettgers (2020), however, that such criteria are in some sense so general that they do not constitute a complete explanation. Likewise, they do not explain the fact that some particular models are transferred while others are not.

Phenomenological criteria, in the form of an ability to reproduce certain patterns, may overlap across domains if the pattern is universal. Universal patterns are abstract structures that can be fitted to facts about phenomena in multiple domains through coupling with domain-specific empirical content. Why we observe such patterns is an ontological question that may tell us something about how nature self-organizes into typical structures, or about our way of responding to the epistemological limitations of grasping nature's complexity.

The case of the Yule process provides us with evidence that universal patterns are what enables model transfer in some instances. The case shows how the Yule distribution shaped the original Yule process model to a large degree. Stripped of its ontological content, the Yule process is a device that generates a specific pattern in an analytically tractable way. The reason why Simon and Bonini (1958) uses the same model to construct a model of firm growth is clear: the model was able to reproduce a specific pattern. It was this phenomenological validation criterion that enabled the model transfer. Importantly, the pattern is the starting point for Simon and Bonini (1958), and not the way in which the mechanisms of the model, proportional growth, and constant addition of new entities could be made to apply to firms instead of genera.

The Yule process case study presents us with an instance in which overlap in phenomenological criteria was the primary reason that the particular model of the Yule process was transferred between domains. It is important to state, however, that in other cases (e.g., Knuuttila and Loettgers (2020)), the primary reason for model transfer may overlap in theoretical and/or mathematical criteria.

Chapter 5

Framework: Model Construction and Validation

This chapter presents a framework that integrates the various insights derived from the cases in the previous chapters. It builds on the existing literature, including Barlas (1996), Kydland and Prescott (1996), Boumans (2007) and Boumans (2009). Importantly, as discussed in Chapter 1, the method through which this framework is derived is primarily inductive. It does not start from any particular philosophical idea about what models are or should be. Rather, it seeks to provide a more systematic understanding of practice in the way that it is actually observed. While the framework has been derived mainly through the analysis of models in macroeconomics, it may also be found to be useful in other disciplines that rely on models.

The framework is an account of model construction and is built around four interrelated concepts: *model purpose*, *invariance*, *model validation*, and *model scope*. I will review these four concepts before explaining how they are involved in the model construction process and how they relate to each other.

The reason why model construction can be related to the concept of validation rests on the assumption that validation criteria are built into the model (Boumans, 1999), as discussed most extensively in Chapter 4. This is because validation criteria are known beforehand and the model is constructed so that it fulfills these criteria. This most often involves a back-and-forth process of trial and error until the right model is found that yields the right balance of criteria fulfillment relative to some purpose. For these reasons, model construction can be understood within the context of validation.

Some parts of this framework may have already been discussed in some of the previous chapters. In general, this framework will discuss these parts in more depth and explore the connections between them. Furthermore, where relevant, I will also connect the framework here to the more general view of models in Section 1.2.

5.1 Model Purpose

The notion of model purpose has a role to play in all three cases discussed in this dissertation. It is the starting point of the model construction process. Importantly, as will be discussed here, the analysis of model purpose yields valuable information about the epistemic aims of the practitioner, as well as the type of model to be constructed.

Scientific models are constructed for a certain purpose (Barlas, 1996). This purpose is a function of the interests and scientific problems of the model builder. In turn, these interests and scientific problems are influenced by scientific paradigms, policy goals, and/or personal considerations. The role of the model builder, therefore, is to construct a model that is in line with this purpose. An example of a purpose that covers most scientific endeavors is that the purpose of the model is to answer a question (Boumans, 2009). Answering a question is often the primary purpose of a model. The subject of these questions is most often a phenomenon; as the main subject of the question, it can be labeled as the model target. To take an example from macroeconomics, the business cycle is a well-known target phenomenon for models. Depending on the question, a model of the business cycle may serve to explain why the cycle occurs or to predict when the next economic crisis will arise.

As discussed in Section 2.4.1, following Boumans (2009), we can distinguish between three types of questions. The first are why questions, the answer to which can be viewed as an explanation (Barlas, 1996). That is, an explanation that relies on making reference to the structure of the real world. The structure of models that may be capable of answering such questions must have an injective representational relationship (characterized as one-to-one; see Section 1.2 for a further discussion) with the real-world system.

The second are how-much questions, the answer to which is, in essence, a numerical value – a measurement. As a result, this type of question does not require any form of explanation. An example are complex models generated by machine learning algorithms. The purpose of such models is to provide an accurate forecast of a variable of interest. Often, however, the model structure generated by the algorithm is so complex that it is unintelligible and does not provide us with an explanation as to why a certain value is forecast. In this way, the structure of models that may be capable of answering how-much questions does not need to have a representative relationship with the real-world system.

The third are how's-that questions, which require an answer that provides some form of understanding of the real-world system structure. In contrast to why questions, however, how's-that questions do not require an answer that makes reference to the structural elements of the real-world system in the way that they are actually observed. Rather, it is sufficient if the answer makes reference to structural elements that, as a whole, behave *as if* the structure of the real-world system. This implies that the model is an account of how possibly the real-world system is structured. The structure of models that are capable of providing such answers thus correlate with the structure of the real world in terms of a wide range of model outputs.

Note that the discussion of model purpose here is in line with the discussion in Section 1.2 on the different types of structural elements from which models can be built. Why questions require the structural elements to be mainly injective representations of theory and/or data. How-much questions only require the model as a whole to be a non-injective representation of only the variable to be measured. We can conceive of highly artifactual structures that would be consistent with this very limited representative requirement. The structure of such models may therefore be artifactual to a large extent. For how's-that questions, the model structure is to be a non-injective representation of a wide range of real-world data characteristics. This somewhat limits the space of possible model structures. In practice, it is often observed that some elements of the model structure are representative of data or theory, while others are artifactual.

5.1.1 Invariance

A further consideration when it comes to providing an answer to a question is the notion of invariance. Invariance was introduced in economics in Haavelmo (1944) and concerns the degree to which model results remain stable over time, place, and/or policy regime. For example, for a model that provides us with an explanation (that is, an answer to a why question), it is desirable for this explanation to hold over time, such that we can apply the understanding given by the model to future questions. This may seem trivial, but it is in fact a crucial challenge in non-experimental science (Boumans, 2007). To sharpen my earlier statement: The primary purpose of a model is to provide an answer to a question that is to some degree invariant – an answer that remains useful even as background conditions change.

In Chapter 3, we discussed how invariance is an important consideration with the development of calibration and estimation approaches to DSGE models.

5.1.2 Secondary Purposes

In addition to providing an invariant answer to a question, there may also be what I will label as secondary purposes. For example, it may be desirable to position one's model within an existing methodological paradigm (in line with the work on scientific programs by Lakatos (1976)). It may also be that a certain scientific paradigm is internalized to such a degree that the model builder is not aware of it, since the model builder is simply constructing a model based on what is believed to be in accordance with best practices.

As an example, most modern DSGE models require the incorporation of new Keynsian macroeconomic theory. It may be that the model builder incorporates this theory purely because it is part of the paradigm, even though the builder knows it is sub-optimal with regard to the model's primary purposes. Alternatively, if the paradigm is internalized to a sufficient degree, the optimal way to achieve the model's primary purpose – to answer a particular macroeconomic question – may be, from the perspective of the model builder, to use new Keynesian theory.

The terms primary and secondary here should not be interpreted as indicative of the degree of importance to the model construction process. This implies that the framework presented here is, in principle, open to instances in which the secondary purposes determine to a larger extent the choices made in the model construction process compared to the primary purposes.

5.2 Model Validation

The second element of this framework is model validation, which can be understood as the assessment of a model's ability to fulfill its intended purpose. In a more practical sense, this is done through assessing the model's ability to fulfill various validation criteria. The model validation criteria are obtained by translating the model's purpose into more concrete and practical points of reference within the context of model construction. This implies that validation criteria are relevant to the model's purpose, meaning that it can be argued that the fulfillment of the criteria also implies that the model is moving toward its intended purpose. This implies that the validation criteria are determined by the model purpose, and that models with different purposes will differ in terms of which validation criteria are relevant.

Throughout this dissertation, we have seen that validation provides an interpretive structure through which many insights can be derived about modeling practices. Chapter 2 mainly focused on phenomenological validation criteria, which allowed us to gain insight into various aspects of agent-based models. In Chapter 3, phenomenological criteria were also seen to play a significant role, as were technical criteria. More specifically, the relationship between model purpose and these technical validation criteria provided insight into the shift from a calibration to an estimation approach in DSGE models. Finally, Chapter 4 used the notion of overlapping validation criteria as a means of understanding model transfer.

In what follows, I will distinguish between theoretical, phenomenological, and technical validation criteria, as well as their relationship to model purpose. Figure 5.1 presents a schematic overview of the relationships between the model purpose, the three types of validation criteria, and the model.

5.2.1 Theoretical Criteria

Theoretical criteria constitute the assessment of whether a model is in line with relevant established theory. Let us first establish what is meant by theory, as it has been defined in multiple ways in relation to economic modeling. For example, Lucas (1980) broadly defines theory as the set of instructions for building the model economy. However, this is not the definition of theory that I have used use throughout this dissertation.

Theory refers to the conceptualizations of the mechanisms associated with the real-world structure. Theoretical criteria are not instructions; they are constraints derived from existing theoretical notions. New Keynesian economic theory, for example, entails price stickiness. A macroeconomic model that is to be built in line with new Keynesian theory should therefore be able to generate price sticky behavior if it is to meet such a theoretical criterion (Galí, 2015).

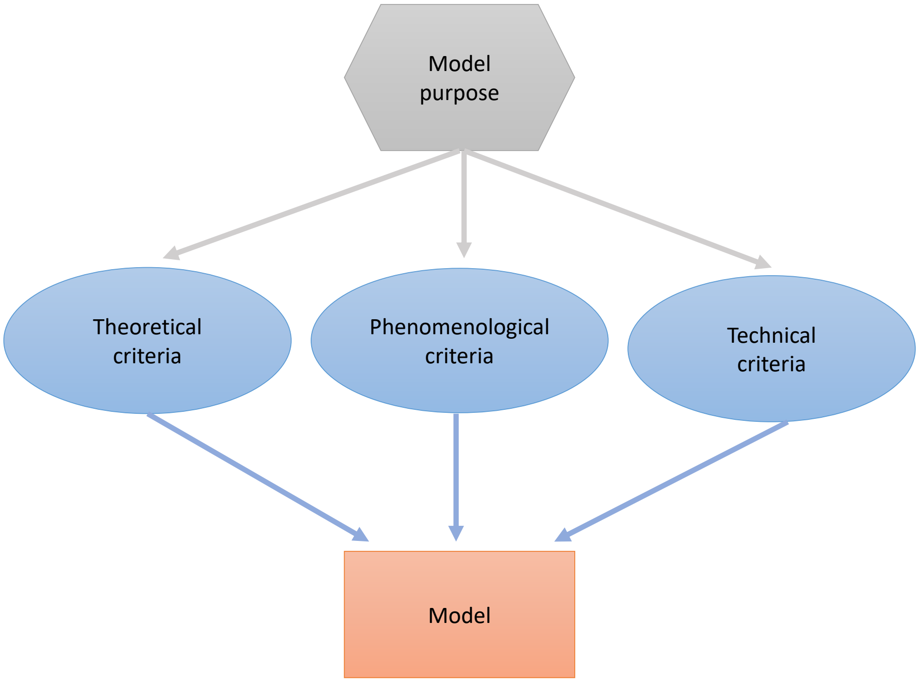


Figure 5.1: Validation Criteria

To illustrate this, in Chapter 4, we discussed the role that theoretical criteria played in the construction of the Yule process model in its original formulation. The context of the evolutionary biology theory of the time constrained some of the model assumptions, especially the notion that genera growth occurs through mutation. In this case, theory provided a framework within which Yule constructed his model. Although this theoretical framework did not completely determine all the model's choices, efforts were made to convince the reader that most of the model assumptions did not disagree with evolutionary biology theory.

5.2.2 Phenomenological Criteria

As we have discussed, phenomenological criteria come in the form of what is more generally known as empirical validation, that is, assessing whether certain model characteristics correspond to what is empirically observed. I will first discuss the various types of phenomenological validation criteria, along with which tests are associated with them. In the next section, I will outline a categorization of model types based on the types of phenomenological criteria that apply. These model types, in turn, are associated with particular types of questions that models are built to answer, as discussed in the previous section.

First, we should distinguish between phenomenological input criteria and phenomenological output criteria (see Gatti et al. (2018) for a similar categorization). Input criteria require the empirical assessment of the model assumptions without considering their implications when brought together. That is, they are the raw ingredients of the model before any analytical derivations or simulations have been performed. In the case of macroeconomic models, this often comes in the form of assumptions at the agent level. We could, for example, seek to assess whether consumer saving behavior is in line with how subjects are observed to form expectations in laboratory experiments.

Phenomenological output criteria, on the other hand, require the assessment of the implications of the model assumptions when put together. This may include both the implications of all model assumptions or subsets thereof. The implications of model assumptions can be studied in many different ways. In some cases, the model is solved analytically; relationships within the model structure can be uncovered in this way and subsequently compared to relationships found in empirical data. In other cases, the behavior of the model as a whole is generated through computer simulations. This yields model output data, the characteristics of which can be compared to their potential empirical counterparts.

We can further delineate these phenomenological output criteria. Barlas (1996) distinguishes between three types of validation tests that can be understood as different types of phenomenological output criteria within the context of the framework presented here. The three types of tests are behavior pattern tests, direct structure tests, and indirect structure tests.

Behavior pattern tests assess whether the model is able to reproduce the major patterns exhibited by the real system. We have labeled these major patterns as the model target, which is determined by the purpose of the model. For instance, if the model purpose is to understand how business cycles arise, behavioral pattern tests would assess whether the model is able to reproduce business cycles as they are empirically observed.

Next, let us discuss direct and indirect structure tests. Both types of tests can be understood as phenomenological output criteria directed toward the model structure. The model structure entails the mechanisms by which the model target is generated. Whether and the way in which the model structure is subject to validation is dependent on the type of question that the model is constructed to answer. For instance, for a why question, an explanation is required; this explanation is embedded in the model structure. It implies that for these question types, the model structure is subject to validation. In contrast, how-much questions do not require an explanation. The answer to such a question is a quantitative measurement that is as accurate as possible. The mechanisms through which the model arrives at this measurement are not of primary interest. This implies that the model structure is not subject to validation.

Let us now discuss the difference between direct structure tests and indirect structure tests. The subject of direct structure tests are subsets of the model structure in isolation from the other parts of the model. They often take the form of particular relationships between variables in the model. For example, it may be presumed that a negative direct relationship is present between interest rates and the propensity to consume. A direct structure test would assess whether this direct relationship is, in fact, present in the model as such, in isolation from other relationships. Direct structure tests therefore assess whether the model structure corresponds to the description of the real-world structure (at least in the way it is observed). I also refer to this type of validation test as direct structure validation. To relate our discussion here to that of the elements from which models are constructed in Section 1.2, direct structure validation assesses whether these elements are injective representations.

Phenomenological input criteria can also be seen as a type of direct structure validation. This is because the model assumptions are inputs to the model structure, and because the validation of these assumptions occurs without considering their relation to the other assumptions or their implications for the rest of the model.

The second type are structure-oriented behavioral tests. These tests assess the model structure, but they do so by comparing a broad range of characteristics of the model output data with their empirical counterparts. The idea is that if the model behaves in line with observed facts in a sufficient number of dimensions, this signals that the model mechanism behaves at least as-if a real-world mechanism. If a model were built to provide an understanding of the mechanisms that could generate the business cycle, for instance, the fact that it is also able to reproduce the Philips curve is seen as a sign that the model mechanisms behave as if the real-world structure. I refer to this type of validation test as indirect structure validation.

In Figure 2.4.2, we can see an overview of all the different types of phenomenological validation criteria discussed so far. The first distinction is between validation based on model input and validation based on model output. Output validation can be split into target and structural validation. Structural validation, in turn, comprises direct and indirect structural validation.

5.2.3 Technical Criteria

As discussed in Section 1.2, a useful metaphor for models is to view them as mediating instruments of investigation. Models are tools that help us answer particular scientific questions. Because these questions are about phenomena observed in the world, models incorporate, to some extent, representations of the world. The purpose of the theoretical and phenomenological criteria discussed so far is to assess these representational aspects of models. However, this is not enough to enable the model to answer the question correctly. The ability of models to answer questions also relies on a technical dimension, which is assessed based on technical validation criteria. In the context of macroeconomic modeling, we can distinguish between mathematical and statistical criteria.

In order for the model to provide answers to a question, the model should work mathematically. This could refer to the analytical solvability of the model, as seen in the case of Chapter 4. In this case, if the model could not have been solved analytically, it would be impossible to consider its implications, the mechanisms of such a model could not be studied, and its output could not be compared to empirical data. Mathematical criteria also include analytical tractability. If the model purpose is to provide an explanation, it is implied that such an explanation enhances understanding. In such cases, therefore, the mathematical complexity of the model is limited.

Statistical criteria require constructing the model in such a way that a certain statistical methodology can be applied. One of the reasons statistical methods are employed when taking the model to data is to ensure a degree of invariance. In what follows, I will make this more concrete.

The main way in which statistical methods are applied in the context of modeling is to ensure the invariant parametrization of the model – that is, to assign a quantitative value to the model parameters. Parametrization is a necessary step in most modeling exercises because it provides us with information regarding the "size" and the "sign" of the relationships within the model structure. If we are interested in the effect of an increase in the interest rate on consumption propensity, we would want to know whether the effect is positive or negative, and how sizable the effect will be. Additionally, parametrization is necessary before any model outcome can be compared to empirical data.

Parametrization techniques are statistical in nature, meaning that they make use of empirical data and certain assumptions about the data-generating process of the model and the real world. Importantly, they are specific requirements of the model structure in order to apply specific statistical methods. An example is including a number of stochastic components into the model to account for any differences between the model and the real-world structure, as discussed in Chapter 3.

The main reason that statistical methods are employed in the parametrization process is to ensure some degree of invariance. In order for the model to provide answers to questions that remain stable over time, place, or policy regimes, it is necessary that the model parameters remain stable as well. One example is the statistical concept of significance. In this context, significance is generally understood as a measure of the likelihood that an estimated parameter is actually zero. The reason behind this uncertainty is that parameters are estimated based on a limited sample of data. Hence, it could be that by coincidence (assuming no selection bias), we have sampled data in such a way that our estimate is nonzero, while the “true” population estimate is actually zero. Only if the estimate is different enough from zero, or the amount of data used is sufficient, can we say that the parameter is statistically significant. Significance can thus be understood as a measure of invariance to changes in the sample of data used.

Another example is the calibration methodology that started with the Lucas critique (Lucas, 1976) discussed in Chapter 3. In this methodology, models should be constructed from “taste and technology” parameters. The idea is that such “deep” parameters do not change as a result of policy interventions. The parameters are then usually indirectly calibrated using stylized facts, which have been shown to be relatively invariant over time.

Statistical methods are also applied to provide an “objective” measure of model performance. This is crucial when it is of interest to know whether a particular innovation in the model structure is an improvement on pre-existing model structures. An example is the likelihood score, which requires the model to be parametrized using maximum likelihood methods.

It may be argued that statistical criteria are part of the process by which we take the model to the data, and that it should thus be seen as part of the phenomenological criteria. Conversely, I argue that it is useful to separate these criteria. Phenomenological criteria concern the relationship between the model and facts about phenomena in a direct sense. Does the model correspond with what we observe? The application of a statistical methodology is in some cases a prerequisite for making such an assessment, but it also retains some degree of independence. There are multiple ways in which we can parametrize a model, for example. As discussed in Chapter 3, the estimation and calibration approaches are two examples of distinct statistical strategies. The ability of a model to reproduce relevant facts about phenomena is therefore not fully dependent on the statistical strategy. This is because the statistical strategy is chosen based on other considerations, such as invariance and providing a measure of model performance, as we have discussed above.

5.2.4 Interdependence of Criteria

An additional complicating factor that may be considered is that the ability of a model to fulfill one validation criterion is often not independent from the fulfillment of the other validation criteria. This implies that model construction, in practice, often comes down to a balancing act between the various relevant validation criteria. As an example, tension may arise between the fulfillment of theoretical and mathematical criteria. Theoretical notions may be complex to such a degree that their incorporation into a model structure would cause the model to become analytically unsolvable, or the model could become so complex that it is rendered unintelligible.

In the same way, theoretical and phenomenological criteria may be at odds. The incorporation of certain theoretical notions into a model structure may imply that the model output is not in line with certain facts about phenomena. In some instances, the modeler has to prioritize certain validation criteria. Of course, any balancing or prioritization of validation criteria is again a function of the purpose of the model.

A further complicating factor may be that, in practice, some validation criteria cannot be identified as being purely theoretical, technical, or phenomenological. For example, the theoretical notions that underlie what we could recognize as theoretical validation criteria may themselves be partially based on empirical evidence. In addition, in models in physics in particular, theoretical notions are sometimes tied

Criteria type	Description
Theoretical criteria	Require the model to be in line with existing theoretical concepts
Phenomenological criteria	Require various elements of the model's input, output, target or structure to be in line with what is observed empirically
Technical criteria	Require the model to have favorable mathematical or statistical properties

Table 5.1: Summary of validation criteria

to particular mathematical formulations. Being able to express a theoretical notion with mathematical elegance is seen as support for that theoretical notion. For the purposes of this paper, however, the categorization presented in this framework is useful, and we will nevertheless be able to classify the relevant criteria as being primarily theoretical, technical, or phenomenological. Table 5.1 presents a short summary of the validation criteria discussed.

5.3 Model Scope

The model scope refers to the domain and type of overlap between the real world and the model. The concept was introduced in Chapter 3 but will be discussed in greater depth here. The position of model scope in the context of this larger framework is that some of the validation criteria require a model of a certain scope. In other words, they require the model to overlap with the real world in certain ways to be fulfilled to a sufficient degree. I will now discuss the various types of overlap that we can distinguish and relate the model scope to the validation requirements.

To clarify the concept of model scope, it is useful to introduce the idea of the real-world data-generating process (rwDGP) and the model data-generating process (mDGP; Windrum et al. (2007)). The rwDGP is to be thought of as a process instantiated by the real-world structure. Each variable that we observe can be thought of as being generated by a web of structural factors that determine its value. However, we cannot directly observe the rwDGP. Instead, we seek to gain a level of understanding of how the rwDGP works through the application of various scientific methods, including the construction of models, which yields an mDGP. The mDGP generates data by running simulations. The idea of scope is thus centered on the relationship between the rwDGP and the mDGP.

The relationship between the rwDGP and the mDGP can be thought of as two-dimensional. The first dimension is the relationship expressed in terms of the degree of overlap between the relevant data generated by the rwDGP and the data generated by mDGP. To put it differently, it is the degree to which the mDGP captures the observed data generated by the rwDGP.
The second dimension concerns the relationship in terms of the structures of the processes that generate data. The structure of the rwDGP can overlap with that of the mDGP in three main ways while holding the overlap in terms of the first dimension constant.

First, it may be that there is no overlap; the model structure and the real-world structure operate differently and have no epistemological relationship. That is, studying the model mechanisms yields no understanding of the rwDGP. The second case involves overlap between the mechanisms in the mDGP and the structure of the rwDGP. That is to say, the mechanisms in the mDGP are injective representations of the actual structure of the rwDGP.

In the third case, a different kind of overlap is evident between the mechanisms. In this case, the mechanisms in the mDGP behave as if the mechanisms in the rwDGP in terms of the wide range of output that the mDGP reproduces. This means that the mechanisms are not the same, but for the purpose at hand, they function well enough to be treated as if they were the same. The model structure is a non-injective representation of this wide range of outputs, and it can therefore be described as artifactual to a large degree.

I have and will refer to these two types of model scope as output scope for the first dimension of model scope and structural scope for the second dimension of model scope. Note that the three types of structural scope distinguished above coincide with the three types of questions and phenomenological validation criteria. We have already established that different types of questions require different types of validation criteria to be fulfilled. This fulfillment, in turn, requires the right model scope. Target validation requires the right output scope, direct structure validation requires overlap as an injective representation, and indirect structure validation requires overlap as a non-injective representation. Furthermore, technical and theoretical validation criteria may also place particular requirements on the model scope. This connection is important to keep in mind as we proceed to discuss the calibrated and estimated DSGE approaches.

Figure 5.2 presents a schematic overview of the notion of model scope. The rwDGP produces empirical data, and the mDGP produces model output data. The output scope concerns the overlap between these elements. The rwDGP and the mDGP are the products of a real-world structure and a model structure, respectively. Overlap between these elements determines the structural scope. Note that there is generally no complete overlap in terms of output; the mDGP will not generate all the observed empirical data generated by the rwDGP, nor can all the simulated data be matched to empirical data. This implies that empirical validation is always a matter of degree – how much overlap is there between the simulated and empirical data? In other words, it is an assessment of whether the output scope is sufficient.

At this point, it is useful to distinguish model scope from the concept of model domain discussed in Chapter 2. We have defined model domain as all model input and output that can be interpreted as an empirical claim. As such, the domain is the space that could be used in the empirical validation of the model because it is connected to the subject matter of the model. The type of question, in turn, determines which parts of this domain are relevant. The model scope, on the other hand, refers to how the model actually overlaps with the real world. Ideally, the relevant model domain coincides with the model scope, but this is something that needs testing.

5.3.1 Model Scope and Validation Criteria

Let us now discuss the ways in which several of the validation criteria may impose certain requirements on the model scope. Validation criteria may require a scope with sufficient overlap with the real-world output or structure. The term sufficient here serves to emphasize that overlap is never complete. Whether through our limited understanding of the real-world structure and the accessibility of its output or through the necessarily limited structure of models, the model and the real world can only overlap to a limited degree. This does not mean that a certain validation criterion would no longer be fulfilled in the case of a larger degree of overlap. Rather, validation criteria point to the minimal scope needed to fulfill the criteria.

Let us start by discussing the relationship between phenomenological validation criteria and the model scope. As discussed, phenomenological output criteria can be directed toward the model target; the phenomenon that the model was constructed to reproduce or explain. Not surprisingly, this phenomenon should be part of the

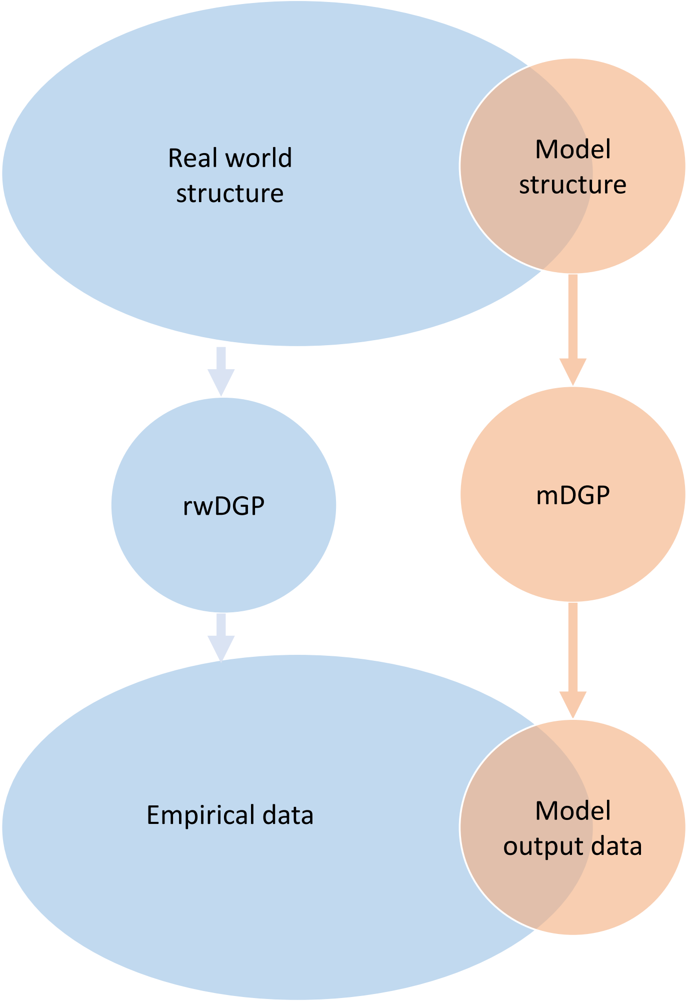


Figure 5.2: Model scope

output scope. That is, the overlap between the simulated and empirical data should include the target, such that the model is able to reproduce it. Additionally, there are the phenomenological output criteria directed at the model structure. These criteria may impose requirements on both the output scope and the structural scope. To comply with direct structure tests, part of the perceived real-world structure should be part of the model scope; to comply with structure-oriented behavior tests, the output scope should include a wide range of empirical facts. A sufficient output scope, in this case, is seen as evidence of at least an as-if type of structural scope. In structure-oriented behavior tests, the structural scope is thus assessed indirectly through the output scope.

Note here that phenomenological criteria serve as a mediator between the model purpose and the model scope. The different requirements of the structural scope ultimately follow from the three types of questions we have distinguished: why, how's-that, and how-much questions. Why questions require that the real-world structure is part of the model scope, how's-that questions require a successful assessment of the structural scope through the output scope, and how-much questions do not require a structural scope. Of course, this does not mean that models constructed to answer how-much questions cannot have a structural scope. A strategy for constructing a model with the purpose of providing measurements may, in fact, be to match the model structure to the real-world structure in some aspects. Given the purpose of the model, the structure is simply not of interest , which is why the model is not required to have structural scope.

What about theoretical criteria and model scope? Recall that theoretical criteria dictate that certain existing theoretical concepts are incorporated into the model. Often, and especially in the case of macroeconomics, these theoretical concepts contain a view of the essential mechanisms in the real-world structure. Or, at least, which mechanisms should be incorporated into the model to have a sufficient as-if relationship with the real-world structure. In this sense, theoretical criteria can be seen as guiding principles for constructing a model that has a sufficient structural scope.

Let us now look at the technical criteria, here grouped into mathematical and statistical criteria. Mathematical criteria require that the model has favorable mathematical properties to ensure analytical tractability, for instance. This relates to the fact that the scope of a model is necessarily limited; that is to say, the mDGP never preserves work by intentionally abstracting, idealizing, and omitting. Models filter out just

	Relation to output scope	**Relation to structural scope**
Theoretical criteria	May require theorized system behavior to be in output scope	Guiding principles to ensure sufficient structural scope
Phenomenological criteria	Target and indirect structural validation requires sufficient output scope	Direct structure validation requires sufficient structural scope
Technical criteria	Certain statistical techniques require sufficient output scope	Mathematical criteria limit the structural scope

Table 5.2: Summary of validation criteria

what is needed to fulfill their validation criteria. From the perspective of mathematical criteria, a pressure thus persists to construct a model with a scope that is no larger than what is needed to sufficiently fulfill the other criteria.

Statistical criteria mainly apply to ensure that the model parameter values are sufficiently invariant. The criteria to ensure the invariance of parameters often require the model to have a certain scope. As an example, in the Cowles Commission's approach to macroeconomic modeling, the statistical strategy included incorporating all essential macroeconomic relations as a means to invariance (Fair, 1992). These relations were not essential for the fulfillment of the phenomenological criteria; instead, they served to fulfill the statistical criterion. To put it differently, invariance required sufficient overlap between the real-world structure and the model structure – a sufficient structural scope. Table 5.2 presents a schematic overview of the pathways between validation criteria and model scope.

5.4 Overview

In this chapter, I have discussed four related concepts that are, in my view, crucial to understanding the model construction process, that is, how a model builder makes choices and manages to construct models. The four concepts are the model purpose, model validation, model scope, and invariance. They are related in the following way: The model construction process starts from a certain purpose. The primary purpose is to provide an invariant answer to a question about a phenomenon; secondary purposes may include positioning one's model within a certain scientific paradigm.

The next step is to translate this purpose into model validation criteria. These validation criteria should be directly applicable to the model construction context. I have highlighted three types of validation criteria: theoretical, phenomenological, and technical criteria. Importantly, validation criteria should be aligned with the model purpose, meaning that the fulfillment of the validation criteria enhances the fulfillment of the model's purpose.

Fulfillment of the validation criteria requires a model of a certain scope, the structure and output of which can be matched to the real world to a sufficient degree. An important note to add is that the validation criteria fully determine the model scope but not the other way around. The reason for this is that the determination of the model structure includes many more choices than the scope of the model alone; there will always be multiple model structures with a similar model scope. The validation criteria determine the full set of structural choices.

Regarding the relationships between these concepts, we can say that the validation criteria are determined by the model purpose. The successful outcome of the model construction process is a model that fulfills the validation criteria. To achieve this, a model of a certain scope is required. The model scope can thus be seen as a requirement that follows from a subset of the validation criteria. Transitively, we can say that the relationship between the required model scope and the model purpose is mediated by the validation criteria.

5.5 Relationship with Models as Mediators

In the introduction of this dissertation, I discussed the account of "Models as Mediators" (Morgan & Morrison, 1999) in Section 1.2. Let us look at the relationship between this account and the model construction framework presented here. Recall that I have distinguished three elements from which models are constructed in the models as mediators account: representations of economic theory, representations of data, and artifactual elements. This characterization is based on the information source. Elements are either informed by theory or data or, in the case of artifactual elements, based on what the practitioner deems necessary for the model to work.

Within the model validation framework presented here, structural elements – or collections of structural elements – are characterized based on the validation criteria that they help to fulfill. Rather than information source, elements are thus characterized based on their function, ultimately in relation to the model purpose.

In this sense, an important question to ask is: What is the relationship between the validation criteria discussed in this chapter and the three information sources of structural elements? For example, do theoretical validation criteria require structural elements to be theoretical representations?

The function of elements within the model is, of course, not independent from its information source. That is to say, in most cases, a strong relationship will exist between the information source of an element and its function. Elements that help to fulfill theoretical validation criteria are, by definition, at least a non-injective representation of that same theory. That is to say, to fulfill the theoretical criteria, the structural choices made must be consistent with the contents of those criteria. This requires at least a relationship of non-injective representation. In the same vein, elements that help the model reproduce certain data are often informed by them. Often, non-injective representations of data are the result of a back-and-forth process between the model structure and its output until the elements are found that ensure the model is able to reproduce the empirically observed output

Since artifactual aspects of structural elements have no representative relationship with either theory or data, their role is often to make the model "work," and as such they have a strong relationship with the technical criteria. If the material from which the model is constructed are mathematical equations, the term "work" can best be understood mathematically and/or statistically. Is the model solvable, or does it yield stable results? These are all what can be understood as favorable mathematical properties.

The model construction framework presented in this chapter can therefore be seen as a reformulation and specification of the models as mediators framework. Instead of information sources, elements are characterized by their functions. This has the advantage of emphasizing why certain representations of theory, data, and artifacts are incorporated into the model structure while others are not. This emphasizes that the model is a product of its purpose and thus helps us understand the choices of model practitioners with this purpose in mind.

Chapter 6

Concluding Remarks

In what follows, I will discuss several more general topics that are relevant to macroeconomic methodology. The aim is not to provide an exhaustive investigation of all these topics. Rather, the discussion offers an overview of the more general picture of macroeconomic modeling that arises from the cases discussed in this study. To some extent, it stretches what can be directly concluded from the cases in this dissertation. It should therefore be interpreted by the reader as an interesting starting point for further methodological discussions.

6.1 Modeling as an Exercise of Integration

To begin, the practice of macroeconomic modeling is an exercise of integration. That is, practitioners actively seek to bring together various distinct elements and transform them into a new entity that functions as one whole. To assemble these various elements, practitioners make use of a particular type of construction material, such as mathematics.

To fit various distinct elements together, they must be adjusted through processes such as idealization. In most instances, this integration requires the inclusion of new artificial elements that have no obvious relationship to data or theory.

Of course, this general view of modeling is not new. As discussed in the introduction, a version thereof is present in Morgan and Morrison (1999) Knuuttila (2021) and Boumans (1999) most notably. Rather, the innovation of this dissertation is a specific view of the elements to be integrated – elements that have a particular role in terms of the fulfillment of validation criteria. Relating model construction as integration to the concept of validation has allowed us to stay particularly close to the modeling practice and has yielded a rich framework that also ties in concepts such as model purpose and model scope.

Within this view of modeling as the integration of validation criteria, several more general statements can be made about macroeconomic modeling practice. First, a direct relationship exists between the primary and secondary purposes of the model and which criteria are to be integrated. In Chapter 2, we have seen that in agent-based models, the model purpose shifts more toward providing a structural account of out-of-equilibrium macroeconomic phenomena. This requires a different set of theoretical and phenomenological criteria at the level of the agent. In the same vein, in Chapter 4, we have seen that the desire to explain a particular observed fact is what primarily determines the validation criteria to be fulfilled.

Second, as we have seen in all the discussed cases, the criteria to integrate are often at odds with each other. In Chapter 3, for instance, I described how the desire to parametrize the model through estimation complicates the structure validation process. Similarly, in Chapter 4, we find an explicit discussion of the necessity to incorporate assumptions that are empirically false but which have to be made in order to solve the model analytically. In Chapter 2, we discussed how, as a consequence of heterogeneity at the agent level, MABMs cannot be solved analytically and thus have to be simulated, which has consequences for the degree to which we can understand the internal structure of the model.

The integration process therefore involves balancing and weighting the validation criteria. In practice, this involves a process of trial and error to gauge how the fulfillment of one criterion affects the fulfillment of another. In this way, the practitioner needs to find the right balance; which balance is correct is again a function of the model purpose.

6.2 One Size Does not Fit All

The second insight follows from the first: no model can do it all in macroeconomics, nor I am confident that we could work toward such a model in the future.

The absence of an all-encompassing model in macroeconomics means that there are limits to the breadth of the model purpose that can be translated into a set of fulfillable validation criteria. That is to say, we can formulate model purposes that are so broad that they lead to a set of validation criteria that cannot be fulfilled within a singular mathematical object: the model.

We can distinguish two ways in which a model's purpose can be broadened. Note that we have characterized the model purpose, in most cases, as answering a question about a phenomenon. First, we see a broadening in terms of the number of phenomena that are relevant to the question. In addition to seeking to explain the dynamics of the business cycle, we could, for example, also seek to explain the dynamics of long-term economic growth. Such a model can found in Dosi et al. (2010), for instance.

Second, we see a broadening in terms of the types of questions to be answered by the model. We have distinguished between why, how-much, and how's-that questions. Typically, the model purpose can be characterized as being primarily one of those three. In some instances, however, multiple types of questions can represent the purpose of a particular model. In Chapter 3, for instance, we have seen that the introduction of the estimation approach in DSGE models was partially motivated by a desire to make quantitative predictions (how much questions) in addition to answering how's-that questions.

The reason that both types of broadening are limited in their feasibility is that, as discussed before, validation criteria are often at odds with each other. Each type of broadening brings with it a particular type of tension between validation criteria.

First, when it comes to broadening the model purpose in terms of the number of phenomena that are relevant, this usually involves increasing the number of economic relations to be taken into account within the model structure. In the case of mathematical models, this generally implies an increase in the mathematical complexity of the model. If this increase in complexity is significant, it weakens the model's ability to fulfill mathematical criteria, such as analytical tractability. As discussed in Chapter 3, this has been a criticism of the models put forward by the Cowles Commission, since their models became so large that they were no longer intelligible (Boumans, 2009).

Second, in the case of broadening the model purpose in terms of the type of question to be answered, a tension is found generally between phenomenological criteria. Recall that models constructed for how-much type questions (black-box models) place a heavy weight on the empirical validation of the model target. Validation of the structure by which the model is able to reproduce the model target is not required. This implies that the model structure can serve purely to optimize the reproduction of the model target. Often, such models are employed when prediction is the primary purpose of the model. In most cases, in fact, this results in models that are not able to provide explanations; they are not able to answer why questions. Their structure is not suitable for this purpose either because it is so complex that it is unintelligible (models generated by machine learning algorithms, for example), or because it incorporates structural elements that do not have an economic interpretation (lagged and stochastic terms).

A model that both seeks to provide answers to how-much and how's-that questions is thus faced with a tension in which either the accuracy of the prediction suffers (target validation) or the ability of the model to provide an empirically valid account of mechanisms (structural validation) does so. This reflects a tension between various phenomenological criteria in that target validation and structural validation are, in some cases, at odds with each other. A primary example of this is discussed in Chapter 3, in which the structure of estimated DSGE models is described as hybrid in this sense.

Nonetheless, the fact that the breadth of models is inherently limited does not mean that progress in this direction cannot be made. When it comes to the mathematical complexity of models, for instance, we have seen the development of new methods for analyzing model results (Guerini & Moneta, 2017), thereby enhancing the analytical tractability of complex models. In a similar vein, a host of methods now exist that aim to open up the black-box of deep neural networks (Montavon, Samek, & Müller, 2018). However, this does not take away from the fact that since our cognitive limitations are ultimately bounded, so too is the extent to which we can increase the complexity of models while still acquiring understanding from them.

What are the implications of this for scientific practice? Above all, it is a call to embrace methodological pluralism when it comes to models in macroeconomics. Particular types of models are best suited for particular types of questions, and limiting the number of phenomena to be explained within one model enhances its intelligibility.

In this dissertation, I have discussed several types of macroeconomic models: agent-based models, calibrated DSGE models, estimated DSGE models, and simultaneous equation models. In addition, in the account of model transfer in Chapter 4, we have discussed more general models of growth processes. Often, as we have seen, these models are presented as if in competition with each other – striving to become the workhorse model of macroeconomics. What can we learn about such disagreements, given the investigations in this dissertation?

Since the aim of this dissertation is to gain a more systematic understanding of existing scientific practice, it is beyond the aim of this dissertation to judge the epistemic value of the various scientific aims and purposes of competing modeling practices. What we do have, however, is a framework that allows us to understand and pinpoint why certain choices are made in the construction of models. This also allows us to explain differences between models of competing paradigms. The added value of the framework presented in this dissertation, therefore, lies in its clarification of the sources of possible disagreements between modeling practices.

Different types of models are best suited to fulfilling a particular set of weighted validation criteria. Given that these validation criteria are a function of the model purpose, this points to differences in the purposes for which models are constructed, or alternatively, in how the model purpose is best translated to the validation criteria. Sources of disagreement over how best to construct models are hence the result of a different view on what the model purpose should be, or a different view on the relationship between the model purpose and the validation criteria.

When it comes to the primary purpose of a model, epistemic disagreements may arise over which phenomena are worth investigating and on which types of questions macroeconomic practice should focus. When it comes to differences between modeling practices that arise from differences in the primary purpose of the model, one could also take a pluralistic view. In such a view, differences in modeling practices can be viewed as something positive: a rich scientific landscape equipped to answer a variety of different types of questions about a large number of economic phenomena. The "one size does not fit all" notion that arises from the investigations in this paper is, in fact, one that is in line with a more pluralistic attitude toward economic methodologies in those cases where the differences in methodologies arise from differences in primary model purposes.

The secondary purposes of models may also be subjected to critique and disagreement. As we have discussed, secondary purposes include, for instance, seeking to fit one's model within a certain existing paradigm. Often, a version of this critique is laid at the feet of DSGE models (Stiglitz, 2011). Rather than seeking to understand the real world, DSGE modelers have focused on the world they created within their modeling structures. This may have led DSGE modelers not to incorporate structural elements in their representative models of, most notably, the financial sector, which turned out to play a crucial role in causing the 2008 financial crisis.

A model may also draw criticism when there is an apparent mismatch between the model purpose and the validation criteria. A critique formulated by agent-based modelers, for example, is that the phenomenological input criteria are non-existent in DSGE models, while their purpose is to provide us with explanations (Farmer & Foley, 2009), meaning that the assumptions about agent behavior in DSGE models – such as the rational expectations of consumers and producers – are not assessed in terms of their empirical accuracy. The behavioral assumptions in agent-based models, as discussed in Chapter 2, are validated using evidence from experimental data. It should be noted, however, that such a critique may also be caused by a misunderstanding or an overstatement of the purpose of DSGE models. As we have discussed in Chapter 5, if the purpose of DSGE models is to answer how's-that or how-much questions, direct structure validation in the form of input criteria is not required.

6.3 Models as Correlational Artifacts

Models are tools for answering questions. They often do so by integrating elements that are representative of relevant theories or empirical data. As I have shown throughout this dissertation, however, this description of models is incomplete. The construction of models necessitates the incorporation of elements that are, at least to some degree, artifactual. Choices must be made that are not entirely based on information that we have about the real-world system. They are choices that have to be made to make the model work – that is, to enable the model to fulfill its purpose. The degree to which a model is constructed based on representational information may differ depending on this purpose. However, the fact is that in all cases, macroeconomic models rely on artifactual elements to a significant degree.

What, then, are models in relation to the real world? How do we learn about the real world from such artifactual structures? One way to understand this is to think of the model as an artifact that has, to some extent, stable correlation with the real world. Correlation here means that, for whatever reason, certain behaviors or patterns tend to co-appear in both the model and the real world. These correlations can be limited to one variable, as is the case in models constructed to answer how much-questions, expanded to a wide range of patterns, as is the case in how's-that questions, or they can be examined in terms of the individual structural elements, as is the case for why questions.

This notion of models as artifacts that correlate with the real world is in line with the account presented in Simon (1969). The well-known example is that of a clock (Hoover, 1995) that tells time. The clock has an internal environment, which is its mechanical structure. This internal environment should be appropriate for its external environment: the mechanism's structure should be such that the clock accurately tells the time. This means that there should be a particular correlation between the internal and external environments. Importantly, the internal environment does not need to be an injective representation of the structure of the external environment to fulfill its purpose. In fact, clocks can work through different internal environments while being able to correlate with the external environment in the same way. Of course, given a different purpose, we may require artifacts to correlate with the external environment along many dimensions. The essence is that the internal environment of artifacts retains some independence from their external environment.

In the case of models in macroeconomics, practitioners seek to establish this stable correlative relationship between the internal (model structure) and external (real world) environments through the fulfillment of various types of validation criteria. The fulfillment of theoretical criteria often requires the model to incorporate certain mechanisms that may help establish a correlation with the real world. We can think, for example, about the commitment to some theoretical notions from evolutionary biology in the Yule process from Chapter 4. Here, evolutionary theory guided the construction of the Yule process model.

Phenomenological criteria seek to establish whether the model output data correlate with the real-world data within relevant domains. If empirical validation tests establish that these correlations exist within the sufficient number of domains, the practitioner may be confident that such correlations also exist for untested domains. This is necessary to analyze, for example, the possible effects of economic policies. In Chapter 2, for instance, we have seen that a large number of empirical facts are compared to the output of agent-based models to establish whether correlation exists in a sufficient number of domains. From this, it is then assumed to be plausible that agent-based models can be used to assess the effects of economic policy.

Technical criteria seek to establish whether these sample correlations are stable across time, policy, or geographical location. This is to prevent the correlation no longer holding true when such contextual factors change. We have seen an example of this in Chapter 3, where the calibration approach of DSGE models requires the model to be constructed from “deep” parameters that are invariant against economic policy interventions.

However, correlation does not necessitate a relationship between the ontological structure of the world and the structure of the model. That is to say, that the model structure isolates the essential economic mechanisms that exist in reality (Mäki, 1992). If this were the case, we can ask which characteristics models should have in relation to the real world that would make such a view of models plausible. In my view, two such characteristics exist: accuracy and high degrees of invariance. Invariance here should be understood in a broad sense, meaning that a correlation is stable over time, across different locations, and under different policy regimes. If we look at the performance of models in physics (at least within experimental setups), both of these characteristics seem to hold true. Models in physics predict well within experiments, and they can be performed many times across different locations and at different times. Moreover, models in physics can be used to predict the effects of various interventions within the experimental setup.

For macroeconomic models, these characteristics do not seem to hold true. First, in terms of accuracy, we know that for macroeconomic models, forecasting comes with high degrees of uncertainty (Aikman et al., 2011). We can think of the inability of macroeconomics to accurately project inflation, which has recently been a focus of discussion (Chahad et al., 2022).

The second characteristic of invariance also does not seem to hold true in the case of macroeconomics, at least not to the degree that we associate with, for instance, models in physics. For one, it has been observed that models may perform well during some periods of time but poorly during others, which implies that the stability of economic models is limited. This forms part of the critique of DSGE models discussed in Chapter 2. DSGE models performed well during the 1990s and early 2000s, during a time known as ”the great moderation“ (Stock & Watson, 2002). This was a period in which the observed variation in most relevant macroeconomic indicators was small. Given that DSGE models rely strongly on equilibrium assumptions, it makes sense that their correlation is strong and stable during times of only small deviations from what were perceived to be equilibrium values. In general, however, DSGE models were unable to capture the large deviations from equilibrium observed in the 2008 financial crisis. For such periods of time, agent-based models (or strongly adjusted DSGE models) were better able to correlate with the observed dynamics.

The same is true for invariance to policy regimes, which, according to an interventionist account, is the definition of causation (Woodward, 2005). To illustrate, in addition to not being able to predict or reproduce the dynamics of the financial crisis, DSGE models were also unable to predict the effects of economic policies during this period (Stiglitz, 2011). Whether DSGE models can be used for policy analysis is an active source of debate among macroeconomists. We have seen that one of the main points of Chari et al. (2009), as discussed in Chapter 3, was that if a model relies on stochastic terms to a large degree, it cannot be used for policy because such terms are not invariant against policy.

In terms of their accuracy and invariance, macroeconomic models should not be characterized as isolations of the essential economic mechanisms. Rather, models are systems that can be constructed so that they behave as if the real-world system within the context of some limited window of time, particular geographical locations, or policy regimes, and with significant degrees of uncertainty. The question of how well macroeconomic models can be used to predict over time or the effects of policy depends, therefore, on *how limited* the accuracy and stability of models are seen to be.

The notion of limits to the accuracy and stability of the correlational relationship between macroeconomic models and the world is not, in my view, a reason to state that macroeconomics is somehow underdeveloped as a science. It is important to realize that the scientific cases in which correlations appear very strong and stable are within the context of laboratory experiments (Cartwright, 1999). As argued in Chapter 1, however, macroeconomics is not a laboratory science. Economies are complex systems consisting of many moving and interacting parts. In this respect, they are similar to models in climate science. In contrast to these models, however, the fundamental units in macroeconomics are human individuals, groups, and institutions, each of which represents a complex system in their own right. The macroeconomy is thus a complex system, the parts of which are also complex systems.

One may argue, of course, that macroeconomics will become more accurate and invariant over time because the amount of data we have about the macroeconomy also increases over time. Data are generated by the macroeconomic system within a variable context. Data collected at different points in time, in different geographical locations, or under different policy regimes are thus generated given varying contexts.

The more data are collected, therefore, the better we are able to construct models that correlate with the real-world system in a larger variety of contexts, thereby enhancing invariance and accuracy. That said, this is contingent on the degree to which the real-world structure that generated observations in the past is comparable to the structure in the future. To some degree, we can learn from economic events in the past to guide our actions in the future. The economic structure of the past will generally be comparable to that of the future in some aspects. However, given the complexity of the economic system, coupled with rapid and novel developments in factors that are likely to be of relevance to the real-world economic structure (such as technological innovation, economic policy, geopolitics, demographics, and climate change), the economic structure of the past will likely also be incomparable to that of the future in many aspects.

The lesson for scientific practice here is that model structures should always be open to revision in light of new empirical developments. In the same vein, it is unlikely that we will settle on a particular macroeconomic structure that is useful across widely varying policy regimes, over longer time horizons, and across various geographical locations.

Practice can respond to newly observed economic developments in two crucial ways. First, existing model structures can be altered to fit particular data more effectively. This most often involves the addition of new validation criteria to the existing set, as seen in the wake of the 2008 financial crisis. DSGE models are now generally supplemented with financial frictions to account for the financial sector, which was seen as the main driver of the 2008 financial crisis (see, for example, (Christensen & Dib, 2008)).

The second method is to start from a new model structure. This implies a complete revision of the set of validation criteria associated with a particular model purpose, and potentially the building material as discussed in the introduction. The example studied in Chapter 2 are agent-based models, where the focus is on the reproduction of phenomenological criteria that are largely different from DSGE models as a different building material, namely agent-based computer simulations.

Methodological innovations in these ways are thus a natural consequence of the relatively inherent lack of stability of the correlational relationship between macroeconomic models and the real-world economic structure. Innovation should therefore be encouraged. An attitude in which the main macroeconomic questions are seen as somehow resolved stands in conflict with the nature of macroeconomic modeling. Infamously, Blanchard (2009) remarked that the state of macro was good and that the "central problem of depression-prevention has been solved." This was written shortly before the inability of macroeconomic modeling to be of aid during the 2008 financial crisis became clear.

6.4 Empirical Validation, Realism and Instrumentalism

The above discussion points toward an instrumentalist view of models in economics. In other words, models are useful instruments for providing understanding or for predicting, but that they do not necessarily inform us about the structure's unobserved reality. This is in contrast to a realist view of economic models, which would entail that models inform us about the essential economic process as they exist in reality. These two concepts present us with a dichotomy of epistemic views.

While the instrumentalist view is indeed aligned with a view of models as correlational artifacts, the realism instrumentalist dichotomy is still useful in informing us about the motivations behind a particular empirical validation approach. For instance, it is sometimes claimed that the reason why agent-based models are empirically validated as they are is to bring them more in line with a realist view of modeling (Windrum et al., 2007). The purpose of the discussion here is to demonstrate how the realist instrumentalist dichotomy may play a role in the empirical validation process and to give us a better understanding about empirical validation practices in relation to the realist instrumentalist dichotomy. As an example, I will investigate the aforementioned claim that agent-based models are empirically validated more in line with a realist view compared to DSGE models.

To understand how it is necessary to discuss in more depth how empirical validation enhances the perceived correctness of a model from an epistemological point of view. We have discussed this issue in Section 2.4.4 I will repeat some points from that discussion here.

A model's answer is composed of a target and a structural account of the mechanisms by which said target was generated. This coincides with the notion of an explanation, which consists of an explanandum –that which is to be explained (i.e., the model target) – and the explanans from which the explanandum follows (i.e., the model structure). In principle, we could say that all empirical data that are somehow a reflection of elements in either the explanandum or the explanans are within the domain of the model. We can think of an example of an answer provided by the model where the explanandum is the business cycle and the explanans is an account of the mechanisms that generate this business cycle.

This implies that for elements within the model target or the model structure that are not in line with the available empirical evidence, we know that the answer provided by the model, or at least part of it, is empirically false. If the elements are empirically observed, we know that the model's answer is not empirically false given the available data. Importantly, any set of available empirical data is always underdetermined by the model's answer. That is, in principle, there are always multiple ways in which a model can reproduce any set of data (Stanford, 2009). We can say that for a given set of data that is within the domain of the answer of model A, we can always conjure a model B that is also consistent with this same set of empirical data. The epistemic value of empirical validation is therefore not to definitively establish that the model is empirically true, but to make it less likely that it is false. In this sense, empirical validation serves to reduce underdetermination.

Given this epistemic value of phenomenological validation criteria, note that the more relevant empirical facts are involved in the validation process, the more underdetermination can be reduced as long as these facts are within the domain of the model's answer. This is in line with the view of some modeling practitioners that models that are able to reproduce a larger number of empirical facts are more epistemically valued.

What is the relationship between empirical validation as reduction of underdetermination and the realist instrumentalist dichotomy? In order to understand this, we should consider that the subsets of the model domain that should be in line with empirical evidence to enhance model validity may be different for an answer provided by a model in line with an instrumentalist view compared to a realist view.

To understand this difference, we should recall the discussion of model purpose in Chapter 5. Modeling within a realist view presumes that the relationships within the model structure ought to exist in reality. This implies that the answer provided by the model is an explanation that involves an ontological description of reality. The purpose of a model within a realist view is therefore to provide us with such an explanation. This can be understood as a stricter specification of why questions as discussed in the previous chapter.

Within an instrumentalist view, scientific methods such as models are not considered capable of informing us about the ontological structure of reality. The purpose of models is therefore not to provide explanations of the kind described in the previous paragraph. The purpose of modeling within an instrumentalist view is, in some sense, more open. Generally speaking, this purpose is strongly connected to the function of the model – that is, what the practitioner aims to do with the model. In this regard, we have distinguished two types of model purposes: to answer how-much questions and to answer how's-that questions.

In Section 2.4.3, we discussed how these types of questions are associated with different types of models and different types of structure validation. Why questions are associated with white-box models and direct structure validation, how's-that questions with gray-box models and indirect structure validation, and how-much questions with black-box models and no structure validation (see also Table 2.1).

The above categorization implies that a more realist view of modeling requires direct structure validation, meaning that the individual relationships present in the model structure are subject to empirical validation. This means that all information within the model domain that is in line with empirical evidence enhances the validity of the model within a realist view. Within an instrumentalist view of modeling, which information within the model domain is subject to validation is dependent on the purpose of the model. For how-much questions, agreement between the structure of the model and empirical evidence does not enhance the validity of the model. For how's-that questions, the structure is assessed through the output of the model as a whole. The agreement between the individual structural relationships of the model and empirical evidence does not enhance the validity of the model.

To relate DSGE models and MABMs to the realist instrumentalist dichotomy, let us position both types of models within the categorization laid out above. DSGE models are grey-box models, since their workings are mainly validated using impulse response functions. If the model is able to reproduce the correct response to a variety of economic shocks, the model is deemed correct. Grey-box models, generally speaking, have a modular structure, meaning that they are comprised of multiple sub-models, the interaction of which is part of the model structure. In DSGE models, these modules could be seen as consumer and firm behavior, and they are only validated in so far as they are useful for generating correct output at the level of the model as a whole. In the case of DSGE models, therefore, these modules can be considered black boxes, and DSGE models can be characterized as grey boxes built from black boxes. As I have described in Chapter 2, MABMs are grey-box models but, contrary to DSGE models, they are built from directly validated modules, making their modules white boxes. The question is: What do such characterizations tell us about the purpose of the models in relation to the realist instrumentalist dichotomy?

It is helpful to look at this issue in terms of reducing underdetermination within a model's domain, as we have discussed before. Within a purely instrumentalist view, one would seek to reduce the possible models by looking at which model performs better in terms of output generated by the model as a whole. In MABMs, through direct structure validation at the micro level, however, we are eliminating only potential models whose structure, that take us from the micro level to the macro level, would only be correct from an instrumental perspective (in terms of model output). The reason for this that any model that starts from agent behavior and that does not isolate elements of agent behavior that can be fitted to the empirical information we have about agents, can necessarily only be correct in an instrumentalist sense. Or, to put it differently, the ontological structure of the macroeconomy must in some sense be derived from the ontological structure of economic behavior at the agent level.

It is helpful to look at this issue in terms of reducing underdetermination within a model's domain, as we have already discussed. Within a purely instrumentalist view, one would seek to reduce the possible models by considering which model performs better in terms of the output generated by the model as a whole. In MABMs, through direct structure validation at the micro level, however, we are eliminating only potential models whose structure, which takes us from the micro level to the macro level, would only be correct from an instrumental perspective (in terms of model output). The reason for this is that any model that starts from agent behavior – and which does not isolate the elements of agent behavior that can be fitted to the empirical information we have about agents – can necessarily only be correct in an instrumentalist sense. Or, to put it differently, the ontological structure of the macroeconomy must, in some sense, be derived from the ontological structure of economic behavior at the agent level.

As discussed in the previous section, the converse is not necessarily true. Because of the fundamental problem of underdetermination, multiple agent-specifications that fit the information we have about agents will always exist. Therefore, the possibility will always remain that an MABM validated at both the agent and the macro level contains a structure that is only correct from an instrumentalist perspective. Again, put differently, it is possible that within the degrees of freedom we have in modeling agent behavior, we make choices that lead to a model whose structure is not correct from a realist perspective, but which still manages to be correct instrumentally. The exercise of validation at the micro level, however, reduces the space of instrumentalist structural elements. This increases the validity of the model from a realist perspective compared to, for example, DSGE models. Therefore, MABMs represent a shift toward a more realist perspective on how to model economic phenomena.

In this way, the realist instrumentalist dichotomy can help us better understand the general focus in the case of MABMs on the validation of individual agent behavior. The desire for realism in macroeconomic models is an active concern in some modeling practices, and this should be taken into account when seeking to understand such modeling practices. Given our discussion of models as correlational artifacts, it is my view that efforts in favor of realism may not be a productive focus given the inherently complex structure of the macroeconomic system.

Chapter 7

Summary

The main aim of this dissertation is to contribute to the systematic understanding of modeling practices within macroeconomics. As its chosen method, it analyzes three particular cases, which, each in their own way, may be taken to be representative of modeling practices today. Subsequently, the results from the case studies are integrated and evaluated.

The first case is discussed in Chapter 2 and focuses on the empirical validation of macroeconomic agent-based models. It is shown that agent-based models in macroeconomics can be best understood by considering them as complex systems with a multitude of interaction levels. Empirical validation tests, as observed in practice, can be related to these levels of interaction. Furthermore, I distinguish between validation tests directed at the model target and the model structure and consider how these apply to macroeconomic agent-based models. The broader insight from this case study is a categorization of phenomenological validation criteria.

The second case is discussed in Chapter 3 and concentrates on the shift in dynamic stochastic general equilibrium (DSGE) models from being calibrated to being estimated. Estimation is considered preferable to calibration for several reasons, in particular because it allows for the parametrization of models with a large number of parameters. In turn, this makes possible the construction of larger models that incorporate more of the complexity of the real-world economic system. This shift required DSGE models to evolve into what I label as a 'hybrid model structure', which is a model with a representational core supplemented with non-representational stochastic elements. This hybrid model structure has come under fire from several authors. I argue that this critique is warranted if it is understood as a disruption in the relationship between the outcome of an empirical validation test and the validity of the model. This case study thus emphasizes the importance of technical and theoretical validation criteria in addition to phenomenological criteria. Moreover, it introduces the concept of model scope, which refers to the overlap between the model and the real world.

The third case, presented in Chapter 4, discusses interdomain model transfer. The case studied is that of the Yule process, a model originally developed in evolutionary biology but later reused as a model for firm growth. The aim of the chapter is to provide an explanation of why such model transfer may appear. It presents a framework of model transfer in which the various validation criteria distinguished in Chapter 3 play a central role. A model is transferred from an original to a new domain when it is found to be useful in both domains. This is the case when overlap is found between the validation criteria of both domains. Special attention is paid here to overlap between phenomenological criteria, which is enabled through the existence of observed universal patterns. Overlap of this type is an important explanation in the case of the Yule process.

The main result of this dissertation is the model construction framework presented in Chapter 5. In this framework, the insights of the case studies are integrated. The framework is centered on the concepts of model purpose, invariance, model validation, and model scope, as well as how these concepts fit together. The main view that it yields is that models are constructed for a certain purpose. Often, this purpose is to provide an invariant answer to a question. This purpose can be translated into more concrete validation criteria, which come in different types. Fulfilling some of these validation criteria requires a particular model scope. Ultimately, this is a framework that provides a systematic understanding of macroeconomic modeling practice. It is useful both for practitioners in macroeconomics seeking to construct models from a systematic basis as well as philosophers of science seeking to better understand observed modeling practices.

Chapter 8

Nederlandse Samenvatting

Het centrale doel van dit proefschrift is om bij te dragen aan een systematisch begrip van modelleringspraktijken binnen de macro-economie. De methodologie van dit proefschrift bestaat uit analyseren van drie casus die representatief zijn voor modelleren in de huidige praktijk. Deze drie studies vormen de basis voor een integrerende en evaluerende beschouwing in het licht van de onderzoeksvraag.

De eerste casus wordt besproken in hoofdstuk 2 en richt zich op de empirische validatie van macro-economische agent-gebaseerde modellen. Het inzicht dat hieruit naar voren komt, is dat agent-gebaseerde modellen in de macro-economie het best begrepen kunnen worden door ze te beschouwen als complexe systemen met een veelvoud aan interactieniveaus. Empirische validatietests (oftewel fenomenologische validatiecriteria), zoals waargenomen in de praktijk, kunnen worden gerelateerd aan deze interactieniveaus. Daarnaast maak ik onderscheid tussen validatietests gericht op het doel en de structuur van het model en heb ik besproken hoe deze van toepassing zijn op macro-economische agent-gebaseerde modellen. De inzichten vanuit deze casus vormen de basis voor een categorisatieschema van fenomenologische validatiecriteria, waar in latere hoofdstukken vaker naar wordt verwezen.

De tweede casus vormt het onderwerp van hoofdstuk 3 en richt zich op de verschuiving in dynamische stochastische algemeen evenwichtsmodellen (DSAE) van zogeheten kalibratie naar meer formele statistische ramingsmethoden. Deze ramingsmethoden hebben verschillende voordelen ten opzichte van kalibratie, met name dat het de kwantificatie van modellen met een groot aantal parameters mogelijk maakt. Dit, op zijn beurt, maakt de constructie van grotere modellen mogelijk die de complexiteit van het werkelijke economische systeem in een grotere mate kunnen bevatten.

De verschuiving vereiste dat DSAE-modellen evolueerden naar wat ik aanduid als een hybride modelstructuur, een model met een representatieve kern aangevuld met niet-representatieve stochastische elementen. Deze hybride modelstructuur is bekritiseerd door verschillende auteurs. Ik betoog dat deze kritiek gerechtvaardigd is als deze begrepen wordt als een verstoring in de relatie tussen de uitkomst van een empirische validatietest en de validiteit van het model. De casus werpt zo licht op het belang van technische en theoretische validatiecriteria naast fenomenologische criteria om een modeleringspraktijk volledig te kunnen duiden. Daarnaast wordt het concept van modelbereik geintroduceerd, dat betrekking heeft op de overlap tussen het model en de echte wereld.

De derde casus in hoofdstuk 4, behandelt het fenomeen van modeloverdracht tussen wetenschappelijke domeinen. De casus betrof het Yule-proces, dat oorspronkelijk werd ontwikkeld als een model in de evolutionaire biologie, maar later werd hergebruikt als een model voor bedrijfsgroei. Het doel van het hoofdstuk is om een verklaring te geven waarom dergelijke modeloverdracht kan plaatsvinden. Er wordt een raamwerk van modeloverdracht gepresenteerd waarin de verschillende validatiecriteria, zoals onderscheiden in hoofdstuk 3, een centrale rol spelen. Een model wordt overgedragen van het oorspronkelijke naar een nieuw domein wanneer het nuttig wordt bevonden in deze beide domeinen. Dit is het geval wanneer er overlap is tussen de validatiecriteria van beide domeinen. In het hoofdstuk wordt in het bijzonder aandacht besteed aan de overlap van fenomenologische criteria, wat mogelijk wordt gemaakt doordat er universele patronen worden waargenomen. Deze vorm van overlap is een belangrijke verklaring voor de modeloverdracht van het Yule-proces.

Het belangrijkste resultaat van dit proefschrift is het modelconstructieraamwerk dat wordt gepresenteerd in hoofdstuk 5. In dit raamwerk zijn de inzichten van de casus geïntegreerd. Het raamwerk draait om de concepten van modeldoel, invariantie, modelvalidatie en modelbereik, en hoe deze concepten op elkaar aansluiten. Het belangrijkste inzicht dat het oplevert, is dat modellen worden geconstrueerd met een bepaald doel. Vaak is dit doel om een invariant antwoord te geven op een vraag. Dit doel kan worden vertaald naar meer concrete validatiecriteria, die verschillende typen kunnen hebben. Het vervullen van sommige van deze validatiecriteria vereist een specifiek modelbereik. Het raamwerk vergroot het begrip van macroeconomische modellering in de praktijk. Het is nuttig voor de bouwers van macroeconomische modellen, die dit willen doen op een meer systematische basis, en voor wetenschapsfilosofen die modelleringspraktijken beter willen begrijpen.

References

Abbasi, A., Hossain, L., & Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, *6*(3), 403–412.

Aikman, D., Barrett, P., Kapadia, S., King, M., Proudman, J., Taylor, T., ... Yates, T. (2011). Uncertainty in macroeconomic policy-making: art or science? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1956), 4798–4817.

Álvarez, L. J., & Burriel, P. (2010). Is a calvo price setting model consistent with micro price data?

Arthur, W. B. (2013). *Complexity economics.* Oxford University Press.

Aydin, A. D., & Cavdar, S. C. (2015). Comparison of prediction performances of artificial neural network (ann) and vector autoregressive (var) models by using the macroeconomic variables of gold prices, borsa istanbul (bist) 100 index and us dollar-turkish lira (usd/try) exchange rates. *Procedia Economics and Finance*, *30*, 3–14.

Baas, N. A., & Emmeche, C. (1997). On emergence and explanation. *Intellectica*, *25*(2), 67–83.

Bacaër, N. (2011). *A short history of mathematical population dynamics.* Springer Science & Business Media.

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review: The Journal of the System Dynamics Society*, *12*(3), 183–210.

Barsky, R. B., & Kilian, L. (2000). *A monetary explanation of the great stagflation of the 1970s.* National Bureau of Economic Research Cambridge, Mass., USA.

Batterman, R. W. (2000). Multiple realizability and universality. *The British Journal for the Philosophy of Science*, *51*(1), 115–145.

Batterman, R. W., & Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, *81*(3), 349–376.

Blanchard, O. (2009). The state of macro. *Annu. Rev. Econ.*, *1*(1), 209–228.

Boumans, M. (1999). Built-in justification. In M. S. Morgan & M. Morrison (Eds.), *Models as mediators: Perspectives on natural and social science* (p. 66-96). Cambridge University Press.

Boumans, M. (2005). *How economists model the world into numbers.* Routledge.

Boumans, M. (2006). The difference between answering a 'why'question and answering a 'how much'question. In *Simulation* (pp. 107–124). Springer.

Boumans, M. (2007). Invariance and calibration. *Measurement in economics: A handbook*, 231–248.

Boumans, M. (2009). Understanding in economics: Gray-box models. In H. W. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding: Philosophical perspectives* (pp. 210–229). University of Pittsburgh Press.

Boumans, M., & Leonelli, S. (2013). Introduction: On the philosophy of science in practice. *Journal for General Philosophy of Science*, *44*(2), 259–261.

Brock, W. A. (1999). Scaling in economics: a reader's guide. *Industrial and Corporate change*, *8*(3), 409–446.

Caiani, A., Godin, A., Caverzasi, E., Gallegati, M., Kinsella, S., & Stiglitz, J. E. (2016). Agent based-stock flow consistent macroeconomics: Towards a benchmark model. *Journal of Economic Dynamics and Control*, *69*, 375–408.

Calvo, G. A. (1983). Staggered prices in a utility-maximizing framework. *Journal of monetary Economics*, *12*(3), 383–398.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science.* Cambridge University Press.

Chahad, M., Hofmann-Drahonsky, A.-C., Page, A., Tirpák, M., Meunier, B., et al. (2022). What explains recent errors in the inflation projections of eurosystem and ecb staff? *Economic Bulletin Boxes*, *3*.

Chari, V. V., Kehoe, P. J., & McGrattan, E. R. (2009). New keynesian models: not yet useful for policy analysis. *American Economic Journal: Macroeconomics*, *1*(1), 242–66.

Christensen, I., & Dib, A. (2008). The financial accelerator in an estimated new keynesian model. *Review of economic dynamics*, *11*(1), 155–178.

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (2005). Nominal rigidities and the dynamic effects of a shock to monetary policy. *Journal of political Economy*, *113*(1), 1–45.

Christiano, L. J., Eichenbaum, M. S., & Trabandt, M. (2018). On dsge models. *Journal of Economic Perspectives*, *32*(3), 113–40.

Corominas-Murtra, B., & Solé, R. V. (2010). Universality of Zipf's law. *Physical Review E*, *82*(1), 011102.

Dawid, H., & Gatti, D. D. (2018). Agent-based macroeconomics. In C. Hommes &

B. LeBaron (Eds.), *Handbook of computational economics* (Vol. 4, pp. 63–156). Elsevier.

De Grauwe, P. (2012). *Lectures on behavioral macroeconomics.* Princeton University Press.

Donhauser, J. (2020). Informative ecological models without ecological forces. *Synthese*, *197*(6), 2721–2743.

Dosi, G., Fagiolo, G., & Roventini, A. (2010). Schumpeter meeting keynes: A policy-friendly model of endogenous growth and business cycles. *Journal of Economic Dynamics and Control*, *34*(9), 1748–1767.

Dosi, G., & Nelson, R. R. (1994). An introduction to evolutionary theories in economics. *Journal of evolutionary economics*, *4*(3), 153–172.

Duarte, P. G. (2012). Not going away? microfoundations in the making of a new consensus in macroeconomics. In *Microfoundations reconsidered.* Edward Elgar Publishing.

Dusenberry, J. S., Fromm, G., Klein, L. R., & Kuh, E. (1965). The brookings quarterly econometric model of the united states. *Chicago: Rand M c Nally Co*.

Edwards, A. (2001). George Udny Yule. In *Statisticians of the centuries* (pp. 292–294). Springer.

Elsenbroich, C. (2012). Explanation in agent-based modelling: Functions, causality or mechanisms? *The Journal of Artificial Societies and Social Simulation*, *15*(3).

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60.

Epstein, J. M. (2006). *Generative social science: Studies in agent-based computational modeling.* Princeton University Press.

Fagiolo, G., Guerini, M., Lamperti, F., Moneta, A., & Roventini, A. (2019). Validation of agent-based models in economics and finance. In C. Beisbart & N. Saam (Eds.), *Computer simulation validation* (pp. 763–787). Springer.

Fagiolo, G., Moneta, A., & Windrum, P. (2005). Empirical validation of agent-based models. In *Acepol 2005-international workshop on'agent-based models for economic policy design.*

Fagiolo, G., Moneta, A., & Windrum, P. (2007). A critical guide to empirical validation of agent-based models in economics: Methodologies, procedures, and open problems. *Computational Economics*, *30*(3), 195–226.

Fagiolo, G., & Roventini, A. (2017). Macroeconomic policy in dsge and agent-based models redux: New developments and challenges ahead. *Journal of Artificial Societies and Social Simulation*, *20*(1), 1. Retrieved from http://

jasss.soc.surrey.ac.uk/20/1/1.html

Fair, R. C. (1992). The cowles commission approach, real business cycle theories, and new-keynesian economics. In M. T. Belongia & M. R. Garfinkel (Eds.), *The business cycle: Theories and evidence* (pp. 133–157). Springer.

Farmer, J. D., & Foley, D. (2009). The economy needs agent-based modelling. *Nature*, *460*(7256), 685–686.

Fernández-Villaverde, J., & Guerrón-Quintana, P. A. (2021). Estimating dsge models: Recent advances and future challenges. *Annual Review of Economics*, *13*.

Gabaix, X. (2011). The granular origins of aggregate fluctuations. *Econometrica*, *79*(3), 733–772.

Galí, J. (2015). *Monetary policy, inflation, and the business cycle: an introduction to the new keynesian framework and its applications*. Princeton University Press.

Gandolfo, G. (2008). Giuseppe Palomba and the Lotka-Volterra equations. *Rendiconti Lincei*, *19*(4), 347–357.

Gatti, D. D., Desiderio, S., Gaffeo, E., Cirillo, P., & Gallegati, M. (2011). *Macroeconomics from the bottom-up* (Vol. 1). Springer Science & Business Media.

Gatti, D. D., Fagiolo, G., Gallegati, M., Richiardi, M., & Russo, A. (2018). *Agent-based models in economics: A toolkit*. Cambridge University Press.

Geweke, J., et al. (1999). Computational experiments and reality. *University of Minnesota and Federal Reserve Bank of Minneapolis*.

Gibrat, R. (1931). *Les inégalites économiques*. Sirey.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Grüne-Yanoff, T. (2009). The explanatory potential of artificial societies. *Synthese*, *169*(3), 539–555.

Guerini, M., & Moneta, A. (2017). A method for agent-based models validation. *Journal of Economic Dynamics and Control*, *82*, 125–141.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, *12(3-4, Suppl.)*, S1-S115.

Hendry, D. F. (2020). A short history of macro-econometric modelling. *Journal of Banking, Finance and Sustainable Development*, *1*(1).

Hesse, M. (1966). *Models and analogies in science*. University of Notre Dame Press.

Holt, C. C. (2004). Forecasting seasonals and trends by exponentially weighted moving averages. *International journal of forecasting*, *20*(1), 5–10.

Hommes, C. (2011). The heterogeneous expectations hypothesis: Some evidence from the lab. *Journal of Economic dynamics and control*, *35*(1), 1–24.

Hoover, K. D. (1995). Facts and artifacts: calibration and the empirical assessment of real-business-cycle models. *Oxford Economic Papers*, 24–44.

Hoover, K. D. (2001). *The methodology of empirical macroeconomics*. Cambridge University Press.

Hoover, K. D. (2015). Reductionism in economics: Intentionality and eschatological justification in the microfoundations of macroeconomics. *Philosophy of Science*, *82*(4), 689–711.

Hoover, K. D. (2021). The struggle for the soul of macroeconomics. *Journal of Economic Methodology*, 1–10.

Humphreys, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*. Oxford University Press.

Humphreys, P. (2019). Knowledge transfer across scientific disciplines. *Studies in History and Philosophy of Science Part A*, *77*, 112–119.

Knuuttila, T. (2021). Epistemic artifacts and the modal dimension of modeling. *European Journal for Philosophy of Science*, *11*(3), 1–18.

Knuuttila, T., & Loettgers, A. (2016). Model templates within and between disciplines: from magnets to gases–and socio-economic systems. *European journal for philosophy of science*, *6*(3), 377–400.

Knuuttila, T., & Loettgers, A. (2020). Magnetized memories: Analogies and templates in model transfer. In *Philosophical perspectives on the engineering approach in biology* (pp. 123–140). Routledge.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (Vol. 111). Chicago University of Chicago Press.

Kurakin, A. (2011). The self-organizing fractal theory as a universal discovery method: the phenomenon of life. *Theoretical Biology and Medical Modelling*, *8*(1), 1–66.

Kydland, F. E., & Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica: Journal of the Econometric Society*, 1345–1370.

Kydland, F. E., & Prescott, E. C. (1996). The computational experiment: An econometric tool. *Journal of economic perspectives*, *10*(1), 69–85.

Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, *3*(1), 33–67.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In S. G. Harding (Ed.), *Can theories be refuted?* (pp. 205–259). Springer.

Lee, F. S. (2012). Heterodox economics and its critics. In *In defense of post-keynesian and heterodox economics* (pp. 120–148). Routledge.

Lengnick, M. (2013). Agent-based macroeconomics: A baseline model. *Journal of*

Economic Behavior & Organization, *86*, 102–120.

Lloyd, E. A. (2015). Model robustness as a confirmatory virtue: The case of climate science. *Studies in History and Philosophy of Science Part A*, *49*, 58–68.

Long Jr, J. B., & Plosser, C. I. (1983). Real business cycles. *Journal of political Economy*, *91*(1), 39–69.

Lucas, R. E. (1976). Econometric policy evaluation: A critique. In K. Brunner & A. Meltzer (Eds.), *The phillips curve and labour markets* (pp. 19–46).

Lucas, R. E. (1980). Methods and problems in business cycle theory. *Journal of Money, Credit and banking*, *12*(4), 696–715.

Lyon, A. (2014). Why are normal distributions normal? *The British Journal for the Philosophy of Science*, *65*(3), 621–649.

Mäki, U. (1992). On the method of isolation in economics. *Poznan Studies in the Philosophy of the Sciences and the Humanities*, *26*(4), 317–351.

Mandelbrot, B. (1982). *The fractal geometry of nature* (Vol. 1). WH freeman New York.

Mandelbrot, B., & Hudson, R. L. (2007). *The misbehavior of markets: A fractal view of financial turbulence.* Basic books.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital signal processing*, *73*, 1–15.

Morgan, M. S., & Morrison, M. (1999). *Models as mediators.* Cambridge University Press Cambridge.

Newman, M. E. (2001). Clustering and preferential attachment in growing networks. *Physical review E*, *64*(2), 025102.

O'Connor, T., & Wong, H. (2015). *Emergent properties (stanford encyclopedia of philosophy.* https://plato.stanford.edu/entries/properties-emergent/. (Retrieved: 2021-04-11)

Palmer, S. E. (1999). *Vision science: Photons to phenomenology.* MIT press.

Parunak, H. V. D., Brueckner, S., & Savit, R. (2004). Universality in multi-agent systems. In *Autonomous agents and multiagent systems, international joint conference on* (Vol. 3, pp. 930–937).

Phelps, E. S. (1967). Phillips curves, expectations of inflation and optimal unemployment over time. *Economica*, 254–281.

Sargent, T. J. (1989). Two models of measurements and the investment accelerator. *Journal of Political Economy*, *97*(2), 251–287.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika*, *42*(3/4), 425–440.

Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society*, *106*, 467–482.

Simon, H. A. (1969). The sciences of the artificial. *Cambridge, MA*.

Simon, H. A. (1976). From substantive to procedural rationality. In J. G. De Gooijer & R. J. Hyndman (Eds.), *25 years of economic theory* (pp. 65–86). Springer.

Simon, H. A., & Bonini, C. P. (1958). The size distribution of business firms. *The American economic review*, 607–617.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, 1–48.

Smets, F., & Wouters, R. (2003). An estimated dynamic stochastic general equilibrium model of the euro area. *Journal of the European economic association*, *1*(5), 1123–1175.

Smets, F., & Wouters, R. (2007). Shocks and frictions in us business cycles: A bayesian dsge approach. *American economic review*, *97*(3), 586–606.

Souleles, N. S. (1999). The response of household consumption to income tax refunds. *American Economic Review*, *89*(4), 947–958.

Stanford, K. (2009). *Underdetermination of scientific theory (stanford encyclopedia of philosophy).* https://plato.stanford.edu/entries/scientific-underdetermination/. (Retrieved: 2021-04-11)

Stiglitz, J. E. (2011). Rethinking macroeconomics: What failed, and how to repair it. *Journal of the European Economic Association*, *9*(4), 591–645.

Stiglitz, J. E. (2018). Where modern macroeconomics went wrong. *Oxford Review of Economic Policy*, *34*(1-2), 70–106.

Stock, J. H., & Watson, M. W. (2002). Has the business cycle changed and why? *NBER macroeconomics annual*, *17*, 159–218.

Suppes, P. (1966). Models of data. In *Studies in logic and the foundations of mathematics* (Vol. 44, pp. 252–261). Elsevier.

Tieleman, S. (2021). Towards a validation methodology for macroeconomic agent-based models. *Computational Economics*, 1–21.

Tobin, J. (1970). Money and income: post hoc ergo propter hoc? *The Quarterly Journal of Economics*, 301–317.

Windrum, P., Fagiolo, G., & Moneta, A. (2007). Empirical validation of agent-based models: Alternatives and prospects. *Journal of Artificial Societies and Social Simulation*, *10*(2), 8.

Woodford, M. (2003). Interest and prices.

Woodford, M. (2009). Information-constrained state-dependent pricing. *Journal of Monetary Economics*, *56*, S100–S124.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Woody, A. I. (2014). Chemistry's periodic law: Rethinking representation and explanation after the turn to practice. In L. Soler, S. Zwart, M. Lynch, & V. Israel-Jost (Eds.), *Science after the practice turn in the philosophy, history, and social studies of science* (pp. 131–158). Routledge.

Yule, G. U. (1902). Mendel's laws and their probable relations to intra-racial heredity (continued). *New Phytologist*, *1*(10), 222–238.

Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. JC Willis. *Philosophical transactions of the Royal Society of London. Series B*, *213*(402), 21–87.