Theory Building for Causal Inference: EITM Research Projects

Thomas Bräuninger and Tilko Swalve

July 14, 2019

forthcoming in: *Luigi Curini and Robert Franzese, The SAGE Handbook of Research Methods in Political Science and International Relations, London: Sage, 2019*

The golden rule that any empirics without theory is, at best, descriptive is likely to be taught to any undergraduate student in her very first year. And there is nothing wrong with this. At the most basic level, the rule reminds us that whatever the co-occurrence of two events, or however strong the correlation of two variables might be, we should not stumble into a fallacy, for example believing that babies are delivered by storks when observing an association between the number of storks and the number of human births. At a more specific level, the rule includes a warning, namely that any sort of evidence for the specific causal mechanism a theoretical argument postulates does not rule out alternative explanations. Data may provide strong empirical evidence for a causal mechanism; at other times, the argument fits into a broader, more general or established theory. Neither case rules out the appropriateness of any other conceivable mechanism in some hypothetical theoretical account. Practically, the question then becomes how to find a method for inference that lets the data speak to the specific mechanism we have in mind, not to any conceivable mechanism. This is the aim of the

Empirical Implications of Theoretical Models (EITM) approach. The EITM approach is a device to 'think about causal inference in service of causal reasoning' (Aldrich et al., 2008).

We should stress that, while the EITM approach seeks to bridge the gap between theoretical and empirical work, it is not just a missing piece to fill a gap in a research design. We think of the EITM approach more as a way to think about related questions. For one, how to improve your methodological reasoning so that empirical work is most effective and informative about theories? Second, how to improve your theoretical reasoning to provide a larger number of useful theoretical hypotheses that can be evaluated against the evidence of empirical models? In doing so, the EITM approach provides a 'coherent approach to evaluating information and converting it into useful and effective knowledge' (Aldrich et al., 2008, 828).

This chapter introduces the EITM approach; discusses its methodological foundations, virtues and challenges; and exemplifies good practices using recent academic work on topical research issues. We discuss three different research designs for how to assess the usefulness of a theoretical model to explore a causal mechanism using data: 1) equilibrium point predictions, 2) comparative statics (the evaluation of relationship predictions), and 3) structural estimation approaches. We introduce each approach with a simple example and pieces of recent research that are exemplars of how to exploit the strength of each design. We finally give some advice on how to strengthen research designs using the EITM methodology.

What is EITM?

The EITM approach promotes the idea that, in the best case, there is a tight connection between a theoretical argument represented by a formal model and an empirical data analysis used to learn when and how the model finds empirical support (or not). There are ontological and methodological presumptions underlying this idea. One is that an ultimate goal of social science inquiry is the explanation of empirical events in the sense of deducing the specific event or statement from a general, theoretical statements about causal relationships. The creation and evaluation of such theoretical statements, then, is an integral part of scientific inquiry. Another assumption is that theoretical statement about states, events and causal relationships can reasonably be captured by formal models.

In making such a claim, we also assume that empirical data can be useful to identify or evaluate the model (neither are the two the same, nor is empirical analysis the only way to evaluate a model; see more on this below), and that evaluation and model identification can tell us something about the usefulness of the theoretical model, and therefore, give us a better understanding of the world. Note that we shy away from using the term 'model testing' here. As Clarke & Primo (2007, 749) remark, models are representational objects; they cannot be true or false. Also, as models serve more than just predictive purposes (they may also be foundational, structural, generative, explicative), they 'should be assessed for their usefulness for a particular purpose, and not solely for the accuracy of their deductive predictions.' [1]

Starting from these assumptions, the diagnosis and motivation for the EITM approach is the observation that much research squanders a great deal of the enormous potential of close links between theoretical and empirical work. This criticism concerns at least one of two points.

First, much theoretical work remains at the level of theory, forgoing the chance to learn more about the *theoretical* argument by exploring empirical data. When the EITM project began in 2001, with an NSF workshop to discuss avenues for improving technical-analytical proficiency in political science, a widely shared concern was the perceived uncoupling of theory building and empirical research as a result of the fast advancement of the discipline's research methods. Since the 1960s, theory building in political science has largely benefited from and partly moved to represent and quantify abstract concepts mathematically, using tools and concepts of social choice, game theory, and microeconomics. This fostered precision in theory building, and hence improved research transparency, lent credibility to this type of work, and resulted in knowledge accumulation. At the same time, the evolution of novel statistical and computational methods fostered the specialization of researchers employing applied statistics and empirical modeling. It was felt that a split had developed between the two camps, or, at least, that graduate training was too often one-sided and created sophistication in either formal or empirical modeling, but not both (National Science Foundation, 2002, 1). The critical question then became how to link rigorous theoretical reasoning and appropriate identification strategies to learn about a causal mechanism.

Secondly, data analysis is often not tied to the theoretical model and, therefore, it is not informative about the causal relationships postulated by the model. Moreover, the focus of analysis often is on one specific causal effect. Much more could be done with a theoretical model, however, in order to derive empirical implications that can then be combined with data to support modeling assumption, provide a better understanding of the causal mechanism, or gain inside into an entire chain of postulated causal relationships. All these types of analyses in the EITM approach offer valuable payoffs.

From Theoretical Model to Implications

We use the terms 'theory' and 'theoretical model' in a quite pragmatic way. A theoretical model is a representation of a causal mechanism. A theory is a set of models that is linked by hypotheses to a set of related features of the real world (Giere, 2010, 85). In this definition and in accordance with common usage in the social sciences, the term 'theory' means a larger structure that collects models that may address different aspects of the real world such as the 'theory of coalition formation' or 'bargaining theory.' The really important term here is 'representation.' Neither does a theoretical model mirror the world on a smaller scale, nor does it simply represent a part of the world. A model simplifies and, in doing so, it focuses more on some aspects and less so on others.

To be sure, there are decidedly different ideas on how the term model should be defined and, consequently, what a model really captures. One view is that there is a *data generating process* that is behind what we observe (Morton, 2003, 33). This mechanisms can be detected, at least in principle. The goal of science is, then, to reveal the mechanism, or get as close as possible to it. From this point of view, it makes sense to say that a model is valid (or invalid). Thus, model testing is an instructive scientific strategy. If the model repeatedly fails to pass some benchmark, it is considered falsified by the data, and thus rejected (Popper, 1959). In this classical falsification framework, the test against data is the relevant benchmark, and any such test probes both the model's postulated mechanism and all (often implicitly stated) auxiliary assumptions.

Another, no less extreme, view is that real world processes are so complex that we can hardly hope to come even close to the true data generating mechanism. From this perspective, models are gross simplifications of the world that, given enough data, would always be rejected (Keane & Wolpin, 2007, 1351). The very idea of model testing thus becomes meaningless. What we can say and do, however, is to gauge whether one model fits the data better than a second one, a procedure one may dub model selection. There might even be alternative models, so that each performs best in some contexts, but not others. The real danger here is to mistake the selection of a proposed and stated mechanism with atheoretical, ad hoc curve fitting that might fit the data well, but consistently performs worse out-of-sample and hardly reveals a deep theoretical structure. A model that relates election results to pre-election polls of the days and weeks before is likely to predict the data with stunning accuracy, but is not very enlightening.

Though interesting, the epistemological debate about what is best conceived as a model, is less important for our discussion. The EITM approach is useful in a wide range of contexts and with different epistemic interests (though we lean toward the second perspective in what follows). In some cases, we are interested in the effect of one specific feature of the world, say X, on some outcome Y. Suppose there are other features, Z, that arguably have an effect on Y, but we have good reasons to assume that some separability condition holds and the relation of X and Y is unrelated to Z. Then, if our (novel) theoretical argument is captured by a model that includes X and Y, possibly in some complex way, a comparative static tells us *how* X is related to Y. A 'test' of the model would focus on the expected covariation of X and Y. The focus here is on the correctness of the model prediction: can we see in the data that X is related to Y? Whether, or to what extent, the model is successful in predicting observed outcome Y is not important here.

In other cases, we are interested in the substantial effect of some rule change, institutional reform or novel policy program. Does a 10 percent decrease in the required majority for a cloture cut the average duration of legislation in half? Does a 5 percent increase in child benefits have a larger effect on education equality than a 10 percent reduction in daycare fees? The focus with these types of research questions is on the size of the change in Y given some intervention X. A model useful to answer the latter question would take into consideration, for instance, how child benefits and daycare fees affect consumption and schooling decisions, while also taking into account expectations about the implications of these decisions in the future, and so on. While some mechanisms are better understood, other critical features of the model, like risk-aversion and foresightedness of parents, have to be estimated as model parameters from the data. As the evaluation of a policy change is based on the comparison of effect sizes, a model must not only capture the key mechanism but also fit the data reasonably well. The approach here is closer to the idea of model selection (via estimation of parameters). What we are interested in is also completeness; that is, how much of the systematic variation in the data that the model aims to explain is captured by the model (Mullainathan & Spiess, 2017).[2]

We also mean a formal model that makes use of symbolic notation and mathematical analysis to state assumptions and deductions (Morton, 1999, 33-74). The key value of a formal model is the clarity of its assumptions. Any theoretical argument that proposes a non-trivial link between a cause and an effect builds on a long list of assumptions about who acts and interacts when, how and with whom. It will also make auxiliary assumptions on the set of cases the model applies to, the nature of the (collective) actors, the cognitive capacities of agents, etc. Often these assumptions are seemingly innocent, as they appear to be common sense or trivial. A formal model makes these assumptions explicit.

Suppose we are studying popular protest in authoritarian regimes. While there are plausible arguments on socio-economic conditions favorable for uprisings, a formal model to capture the theoretical argument will not just explicate what motivates and constrains individuals. Instead, it will also make explicit whether the argument operates on the level of the public at large or on the level of individuals; whether, and, if so, how these differ and how individuals solve their collective action problem; and whether these problems are assumed away by auxiliary assumptions.

Auxiliary assumptions are a key part of a model, as no model can possibly capture every aspect of reality (Low & Meghir, 2017, 34). A model captures something essential of the data generating process, but not everything. A model of legislative oversight might consider variance in institutions, legislative majorities and legislators' preferences for one or another form of oversight, while deliberately ignoring the role of a supreme court (McCubbins & Schwartz, 1984). This is not to suggest that supreme courts do not play any role in a legislator's decision-making. It does assume, however, that decisions over police patrols or fire alarms are not substantially affected by long-term considerations. The question is what cannot be left out and what can be left out to ensure the substantive conclusions of the model remain valid?[3]

A helpful term in answering the question is the term 'separability' . Separability can be understood as the invariance of an ordering on a subspace with respect to changes of variables outside the subspace (Blackorby et al., 1998). In fiscal policy literature, for instance, budgetary politics is often considered as either a top-down or a bottom-up process. In countries with a bottom-up process, so the story goes, individual portfolio ministers draft sectoral budgets, which are then aggregated and finally settled in bargaining. Where fiscal institutions allow for a strong finance minister, the budget process is more top-down: the finance minister decides on the total budget, and the allocation to portfolios is determined by intra-cabinet bargaining. In the first case, the total amount to be spent and the allocation are interrelated, and there is strategic interaction between agents. Whatever the ministerial draft looks like, it will have an effect on both the total budget and the allocation. This is different in the second scenario, where bargaining over the allocation happens when the total budget is already fixed.

When modeling intra-ministerial bargaining, we therefore may consider the total budget as a fixed parameter, rather than reflect on a full model that would model both the finance minister's decision and the allocation decision. The total budget can be thought of as a *sufficient statistics*: something that summarizes decisions made outside the model (Low & Meghir, 2017, 34).[4] From the discussion, it should be clear that practically any political science model makes separability assumptions. Whether or not a model has bite (we deliberate avoid terms like right or wrong, see above) largely depends on the appropriateness of the separability assumptions made.

Figure 1 shows how we think about the relationship between model, theoretical implications and features of the real world:

M **is the model with auxiliary assumptions**: features that we assume

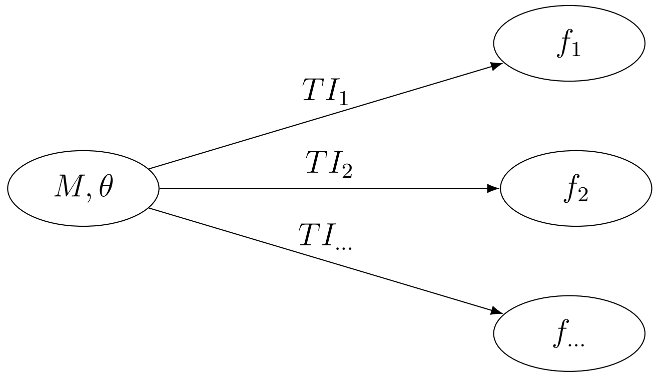

Figure 1: Models, theoretical implications, and features of the real world. Adapted from Clarke & Primo (2012).

about agents and the nature of the interaction: agents' preferences, information they possess, individual information processing, rationality, decision-taking modes, choice sets, strategies available, and equilibrium concept used. We think of these features as being fixed.

θ **is a set of parameters**: features that we think of as fixed in one specific case under study, but that will vary across different cases of interest. Parameters usually serve one of two purposes. Sometimes, we know or suspect that the cases under study vary greatly with respect to some marginal condition such as an individual trait, a sufficient statistic, or institutional feature that we know is critical to the mechanics of the model. Assuming one specific preference relation or utility function, a specific discount factor or voting threshold may amount to having a model that applies only to a single real-world case. If there is reason to believe that we can capture the variation in marginal conditions by some parameter, a parametrized model increases the scope of the model. Comparative statics is, then, a tool to study the effect in the variation of these marginal conditions. At other times, it is unclear whether some marginal condition has any effect on the feature we study. In a parametrized setting, we might want to show that outcomes we are interested in do not vary in the parameter.

$f_1, f_2, \ldots$ **are outcomes** in a very broad sense: features that are associated with the real world and that we argue or predict to occur or vary conditional on M and θ. One common aspect of politics that we are interested in is the actions of agents: Does a foreign state wage a war?; Does the incumbent invest in a costly campaign?; Does the president delegate authority over trade policies to the legislature? This is all observable behavior, even though collecting data on actions taken by agents might be difficult. For instance, while it is pretty straightforward to verify whether a state has declared war on another, collecting data on whether the government has prepared for war might be intricate.

Observable actions must be distinguished from strategies, which are only partially known. Strategies are, using game-theoretical terminology, an agent's complete plan of what action to take at any contingency that might occur. By their very definition, strategies can not simply be inferred from behavior. As they provide a rule for action in response to any contingency, they refer to cases that – in equilibrium – are counterfactuals. Consider the threat to fight. If the threat is credible, it will be successful and deter the opponent from making a certain move, for instance, an attack. To be credible, however, the strategy must include the plan to fight in the counterfactual contingency of an attack. Such a self-commitment is not readily observable.

Another aspect of politics we might be interested in is outcomes, defined as consequences of the equilibrium behavior of agents. These can include policy output and outcomes, the evolution of an institution, some form of coordinated behavior, and an international treaty or war. What we consider an outcome depends both on the puzzle that we wish to understand and on the solution concept we use in our theoretical model.

$TI_1, TI_2, \ldots$ **are theoretical implications** that link all assumptions to features of the real world. Generally speaking, we think of the usefulness of a theoretical model to be weakly increasing with the number of its theoretical implications.

Simple theory, non-obvious implications

Let us consider a simple example of a theoretical model that we seek to combine with empirical data. Working through the example in some detail helps to clarify the steps we have to undertake, the assumptions that need to be made, and the pitfalls we might end up with. Suppose we study the conditions when leaders wage a war. We have an argument that we seek to confront with empirical data. In our simple decision-theoretical model of going to war, a leader has two options. She can accept a previous offer to settle the dispute over a piece of territory yielding utility u_r.[5] Or she can wage a war at costs c, which results in one of two events: a victory generating utility u_w, or a defeat with utility normalized to 0. The odds that the country will win the war is given by some exogenous, fixed and known probability. So, the model is indeed quite simple, as it ignores any strategic interaction with an opponent who himself might prepare for war, etc. Finally, we assume the leader values the two choice options with other criteria that are unrelated to the ones above, ones that we do not know and observe. We assume that disturbances are both unrelated to any other model parameter and known to the leader but unobserved by us (the researcher). For reasons that we will see in a second, we denote these by e_r and e_w.

To evaluate the argument empirically, provided that there is a sufficient number of cases, we would collect data on these cases and how they differ with respect to the above conditions: the gains of winning the war and the gains of settling the dispute without waging a war, but also the costs of a war, and the losses associated with losing war, and so on. As the observed outcome is binary – the leader chooses war or acquiesces – we might run a standard probit or logit model with measures of gains and losses as covariates. Having estimated the statistical model, we can check whether the parameter estimates of the regressors have the expected sign, for instance, whether the costs of war have a negative effect on the occurrence of war, and so on.

However, given our description of model specification, the statistical model is not exactly capturing the mechanism that the theoretical model postulates. The theoretical model suggests that the effect of the gains of winning on waging a war is not unconditional, but instead related to the probability of a victory. Obviously, if the odds of winning are nil, the size of the trophy becomes irrelevant. Thus, any estimation of the model with data that ignores the mechanism will thus not be instructive for the theoretical model. Though this may be obvious here, it may be unapparent with more complex mechanisms and interactions.

In fact, even in this trivial example, walking from the theoretical to an appropriate statistical model involves a number of steps and decisions: decisions that need to be taken and enter the analysis as auxiliary assumptions. To begin with, solving the theoretical model is straightforward here. Let p denote the probability that the leader's state wins. The leader will wage a war if $pu_w - c + e_w > u_r + e_r$ and accept the offer otherwise.[6] With the assumption that the difference in the disturbance has some cumulative density function, F, the probability of waging war is

$$\begin{aligned} q &= \Pr(War|u_w, u_r, p, c) \\ &= \Pr(pu_w - u_r - c > e_r - e_w) = F(pu_w - u_r - c) \end{aligned}$$

In words, the *theoretical model* implies that the probability that a leader wages a war is given by the value of the CDF at the net utility of war.

The nice thing here is that the theoretical model easily translates into a statistical model, as the above equation looks like a standard binary discrete choice model. If we denote the observation of a leader going to war by $y = 1$ and backing down by $y = 0$, we can think of the term $y^* = pu_w - u_r - c$ as the latent preference for war. When the $e_r - e_w$ are independently drawn from a normal distribution with mean zero, variance σ^2 and CDF Φ, the probability that the leaders goes to war becomes

$$Pr(y = 1) = \Phi(\frac{pu_w - u_r - c}{\sigma}).$$

To complete the statistical model, we need to do two more things.[7] First, we need to operationalize the features of the theoretical model, p, u_w, u_r, and c, and then get measures for them. Measurement involves issues of validity and reliability that we do not deal with here. But there is one point we want to stress: suppose that we have a reasonable measure for the leader's *ex ante* probability of winning the war, say π, and also that all relevant costs are material costs that can be estimated ex ante in financial terms, denoted by the measure C. What is less clear is the utility that the leader associates with the two states of winning a war or accepting the deal. Some observable features of the territory – such as size, geography, natural resources and physical capital stock – will be important, but as these goods are not easily convertible, there is no a priori rule as to how exactly different goods jointly determine a utility that is preferences over outcomes.

As a consequence, we have to make auxiliary assumption to fully specify the statistical model. To keep things simple, let us assume the following:

- u_r and u_w are fully determined by two observable features of the territory: its size and its natural resources
- X and Z are valid and reliable measures for size and resources
- all leaders value X and Z in the same way
- the functional form of the relationship between observable features and theoretical concepts is linear, such that $u_r = \alpha_0 + \alpha_1 X + \alpha_2 Z$ and $u_w = \beta_0 + \beta_1 X + \beta_2 Z$ for some unknown parameters α and β.[8]

This is quite a list of assumptions. It is, however, what we do, often implicitly, when estimating models. With the above assumption, we can write the latent preferences as

$$y^* = \pi \times (\beta_0 + \beta_1 X + \beta_2 Z) - (\alpha_0 + \alpha_1 X + \alpha_2 Z) - \gamma C.$$

This is not a linear relationship:

$$y^* = -\alpha_0 - \alpha_1 X - \alpha_2 Z - \gamma C + \beta_0 \pi + \beta_1 \pi \times X + \beta_2 \pi \times Z$$

but involves two interaction terms of π and X or Z, respectively. The important point here is that a simple linear relationship between the regressors and the odds-ratios of going to war does not capture the mechanism behind the choice problem of the leader.

This will even be less so, if there is not a single decision-maker, but two or more actors whose actions and strategies are strategically interdependent. To illustrate, consider a slightly more complex game that considers the strategic interaction of the leader and the opponent state. The opponent state can either prepare for war or not, and either action is observed by the leader. Preparing for war comes at a cost of d for the opponent, but lowers the probability that the leader wins to some $p_0 < p$. While the choice problem of the leader is basically the same – she would account for the opponent's war preparations with p_0 or p – the opponent has a more intricate decision problem. Whether preparing (or not) is preferable largely depends on the probability that the leader wages a war (or not) which again is conditional on the war preparations of the opponent. Using analogous specifications for utility functions v of the opponent, the opponent's choice to prepare is the best strategy if

$$q((1 - p_0)v_w - d) + (1 - q)(v_r - d) + e_p$$

$$> q(1 - p)v_w + (1 - q)v_r + e_n$$

In this case, the statistical model will involve simultaneous estimation of the parameters of the covariates and the choice probabilities. A specification of the statistical model that fails to capture this interaction will not be informative about the causal relationships postulated by the model.

There is a considerable literature on how to derive a statistical model from a theoretical model involving strategic interaction. In the econometric literature, the approach is referred to as structural estimation (Low & Meghir, 2017), though the political science literature is sparse (see, however, Smith, 1999; Signorino, 1999, 2003). The bad news here is that there is no vanilla, or one-size-fits-it-all, method that could readily be applied to very different types of (game-)theoretical models. Decision-theoretical discrete choice problems (as the one above) will often lead to some sort of the statistical random utility model (McFadden, 1974). These may differ in assumptions about the distribution of the unobserved, random utilities (resulting in probit or logit models), and also in the source of the uncertainty they capture (Signorino, 2003): agents may simply err when making choices; they may have private information about their own payoffs; or random utilities represent what is actually omitted variables or measurement error of the covariates that are used to capture features of the utilities.

EITM Research Designs

There is great variety of approaches to combine theory building, model construction and empirical work in political science research. In what follows, we distinguish between three common approaches: 1) equilibrium point predictions, 2) comparative statics, and 3) structural estimation approaches. Throughout, as a running example, we will use a simple entry game (see Figure 2, see Berry & Reiss (2007)), a strategic situation that is ubiquitous in economics and political science, to introduce each research design.

In its original version, two firms decides whether to enter a market (E) or not ($\neg E$). Likewise, we can think of politicians deciding whether to enter an electoral race or not. Entering the market comes at an entry cost c_1 for firm 1 and c_2 for firm 2. If only one firm enters the market, the variable profit is Π^{Mono}. If both firms enter the market, their profits are reduced to Π^{Duo}. If a firm decides not to enter the market, its payoff is 0. We will assume that $\Pi^{Mono} = 2\Pi^{Duo} = \Pi$ and $\Pi > c_1, c_2 > \Pi/2$, which means that each firm will prefer to enter the market if the other firm does not, but it will prefer to stay out if the other firm enters. Players know profits and costs and make their choices simultaneously. The game has three Nash equilibria: two pure-strategy equilibria (E, ¬E) and (¬E, E) and one equilibrium in mixed strategies $(((2 - \frac{2c_2}{\Pi})\text{E},(\frac{2c_2}{\Pi} - 1)\neg\text{E});((2 - \frac{2c_1}{\Pi})\text{E},(\frac{2c_1}{\Pi} - 1)\neg\text{E}))$.

		2	
		E	¬E
1	E	$\Pi^{Duo} - c_1$, $\Pi^{Duo} - c_2$	$\Pi^{Mono} - c_1$, 0
	¬E	0, $\Pi^{Mono} - c_2$	0, 0

Figure 2: Entry Game

Having solved the game, a researcher can confront the theoretical model with data in at least three different ways. First, a researcher can focus on the evaluation of the point predictions of the model. Does the empirical outcome or the employed strategies of real actors match to the equilibrium outcome and the equilibrium strategies? The evaluation of point predictions is often useful in laboratory experiments, but has only limited application in large-N studies relying on observational data. Another important research approach that makes use of equilibrium predictions of formal models is analytical narratives.

Second, a researcher can focus on the relationships between variables that the equilibrium of the game implies. Such comparative statics designs usually do not claim that the formal model they use captures the entire data generating process, but they will seek to evaluate relationships between key parameters of the model empirically. Comparative statics is the dominant research design when combining a formal model with observational data.

Finally, instead of seeking to verify or falsify predictions of the model, a researcher might be interested in structurally estimating parameters of the model, either because they are interesting on their own or because they are useful in making predictions for counterfactual scenarios. We will next consider these approaches in more detail.

Equilibrium Point Predictions

Formal and game-theoretic models are solved by the application of some sort of solution concept, a rule that defines a priori how to single out events that can be expected and, hence, are 'predicted' from the universe of events. For example, models using non-cooperative game theory usually apply the concept of the Nash equilibrium, or some refinements thereof, to make predictions about strategies players use and the resulting outcome of a strategic situation. Models in the cooperative game-theoretical tradition make use of solution concepts such as the core, the Shapley value, or the Walrasian equilibrium that focus on payoff allocations rather than strategies. An appealing characteristic of many solution concepts from a theoretical point of view is that they often make sharp, deterministic predictions.

For example, in the entry game, a sharp equilibrium point prediction can be derived using mixed strategy equilibrium. Let's assume that the total profit is $\Pi = 7$ and firms have differing costs for entering the market $c_1 = 5$ and $c_2 = 4$. Then the mixed strategy equilibrium prediction is that the first firm enters the market with probability $\frac{6}{7}$ and stays out with probability $\frac{1}{7}$, whereas the odds that firm 2 enters or stays out are $\frac{4}{7}$ and $\frac{3}{7}$. These are fairly sharp predictions, but how to use them to evaluate the model?

There are at least four issues. First, as we have shown above, the model has multiple equilibria, firm 1 entering and firm 2 staying out is as much a Nash equilibrium in pure strategies as the reverse case. There is nothing that makes one superior to the other. Second, even if there were only the mixed strategy equilibrium, none of the four observable outcomes could be taken as evidence either for or against the proposed model mechanism. The issue here is that we can only observe behavior, not the actual strategies that individuals use. If both firms indeed use mixed equilibrium strategies, we should observe, for instance, a duopoly with probability $\frac{6}{7} \times \frac{4}{7} = \frac{24}{49}$ and no firm entering the market with probability $\frac{1}{7} \times \frac{3}{7} = \frac{3}{49}$. A straightforward way to evaluate mixed strategy equilibria is thus to focus on these aggregate outcome probabilities, and compare them to relative frequencies calculated from a larger number of observed cases, preferably in a controlled laboratory environment.[9] Third, how to measure Π, c_1 and c_2 in a real market? As we can see above, point predictions of strategies are vastly different when the costs of the two firms are 5 and 4 units, or 4 and 5 units. Fourth, how to make sense of these point predictions if we are not willing to assume that the model reflects the complete data generating process? Even if the model does not miss any systemic feature of the interaction, observed behavior and outcomes will involve some randomness, leaving uncertainty as to their effect and the manner in which to account for it in the model. From an empirical point of view, sharp predictions can be less appealing since the world is generally stochastic and sharp predictions are easily falsified. We are rarely able to study the real world 'in equilibrium.'

How should we assess point predictions? There are three basic routes that one can take (see Morton, 1999). One is to make explicit assumptions about the nature of the random effects on the point predictions that may have prevented those predictions from being observed empirically. This will transform point predictions into distributional predictions. If there is a sufficiently large number of observations, distributional predictions are then assessed against the empirical record. In the market entry game, one could assume that firms perfectly know market profits and costs, but err when making their choice. This results in the random utility framework that we discussed in the section above. In the entry game, taking into account that firms err, the probability p_1 that firm 1 enters the market may be written as

$$p_1 = \Pr(p_2(\frac{\Pi}{2} - c_1) + (1 - p_2)(\Pi - c_1) + \epsilon_{11} > \epsilon_{12})$$

where p_2 is the probability that firm 2 enters the market, and the ϵs are individual and choice specific error terms. Let F again denote the CDF of $\epsilon_{12} - \epsilon_{11}$ and $\epsilon_{22} - \epsilon_{21}$ and the equilibrium predictions are the implicitly defined distributions

$$\begin{aligned} p_1 &= F(-\tfrac{\Pi}{2}p_2 + \Pi - c_1) \\ p_2 &= F(-\tfrac{\Pi}{2}p_1 + \Pi - c_2). \end{aligned}$$

Thus, a solution to the problem – that point predictions are likely to fail to receive support because of a partially modeled data generating process – is to turn the model into something more 'realistic' and/or use an appropriate equilibrium concept, here a quantal response equilibrium (McKelvey & Palfrey, 1995).

A second way is to use controlled laboratory experiments. Laboratory experiments have a number of advantages for evaluating theories, in general, and equilibrium point predictions, in particular. A major advantage of experiments is that they can isolate the theoretical mechanism. The idea is to eliminate as many random effects that occur in the natural environment as possible. The third route is different from the previous ones in that it builds and evaluates a model in the context of a limited number of cases, often a unique case. Historical events such as the French revolution or the rise and fall of Venice may be instructive for our understanding of political mobilization and institutional design, but their study is plagued with the typical problem of case studies: there are different ways to interpret the historical record. Analytic narratives (Bates et al., 1998) is the attempt to provide a causal explanation for such events by building a formal model to capture the logic of the explanation and evaluating it through testable implications. In that sense, analytic narratives is less a method to simply evaluate a model, and more an approach of both model building and evaluation. We present showcases of the latter two approaches in the remainder of this subsection.

Evaluating Equilibrium Point Predictions in the Lab: Battaglini et al. (2010)

An example of the evaluation of the point predictions of a model in a laboratory is Battaglini et al. (2010). Their starting point is the theoretical model that introduced the *Swing voter's curse* (Feddersen & Pesendorfer, 1996). Feddersen and Pesendorfer show that it can be rational for poorly informed voters to abstain in an election, even when the cost of voting is zero. By abstaining in equilibrium, poorly informed voters avoid deciding the election in the wrong direction and, therefore, leave the decision to better informed voters. A central assumption of the model is that voters form beliefs about the probability that their vote will be pivotal. The model matches some empirical observations, such as selective abstention, but the pivotal voter assumption, along with the theory of strategic abstention, remains controversial.

Battaglini et al. use a simplified version of the model. A set of voters decides by majority rule between two alternatives, A and B. Corresponding to the two alternatives, there are two unobserved possible states of the world, A and B. In state of the world A, alternative A is optimal, in state of the world B, alternative B is optimal. There is a number of independent voters (or swing voters) who want to match the alternative and the state of the world, as well as some partisan voters who prefer alternative A or B, irrespective of the state of the world. Each voter may receive a private informative signal about the state of the world, but may also stay uninformed with some probability. After voters have received the signal, votes are cast and an alternative is chosen by majority rule. In equilibrium, all partisan voters vote in favor of their preferred policy, and all independent voters who received an informative signal vote in line with their signal. Uninformed independent voters, however, have incentives to abstain with at least some probability. The reason is that an uninformed independent voter knows that, with some probability, there exists an informed independent voter. By casting a vote, the uninformed independent voter would risk voting against an informed independent voter. The key equilibrium point prediction of the model is, therefore, that independent uninformed voters will abstain with some probability depending on the specific parameterization of the game.

The authors set up a laboratory experiment to evaluate the point predictions of the theoretical model. Values for three different parameters need to be chosen: the number of voters, the number of partisan voters and the prior probability for each state of the world. There are two possible jars (states of the world). The first jar contains six white balls and two red ones, while the second jar contains six white and two yellow balls. A computer randomly chooses a jar with the prior probability. On the computer screen, participants choose one of the eight balls, the color of which is then revealed. This creates a share of informed participants, namely those who see a yellow or red ball, and some uninformed participants, who see a white ball. Finally, each participant decides to either vote for jar A or jar B or to abstain. If the correct jar is chosen by a majority, a higher payoff is paid out compared to the incorrect jar. If there were partisan voters, the participants are told that the computer will cast their votes for jar A and B, respectively. Participants' choices are recorded and compared with the point predictions of the theoretical model.

Battaglini et al. find strong evidence that participants' voting behavior varies as theoretically predicted: depending on the treatment configuration, uninformed subjects abstain, seemingly delegating choice to informed participants.

Analytical Narratives: Gailmard (2017)

In U.S. federal and state constitutions, the separation of power is very pronounced, with governors and assemblies having independent power bases. How did this hallmark of U.S. constitutions evolve? Gailmard (2017) argues that, to understand the origins of the separation of powers in the U.S., we must look at the era of English colonies in North America. What nowadays looks like a natural institution, actually is the result of an institutional choice of a strategic English crown confronted with agency problems in its colonial governors.

Crown-appointed governors in the New World were hard to control from a distance, and, as result, they would over-tax settlers, thereby reducing the incentive for settlers to invest and reducing the revenue to the crown. To help settlers to restrain the governor, the crown created liberal institutions, an empowered assembly with budget power. While this might have negative effects on the crown's revenue in the short run, it should pay off in the long run. By turning over some political control to the colonial settlers – introducing the separation of powers – the Crown could solve the agency problem.

The theory embeds a model of the colonial economy into two alternative political environments, one with hierarchical control and one with a separation of powers within the colony. There are three players, the Crown, a colonial governor, and colonial settlers. Settlers can make either a high or a low investment in the colony, then the governor and the Crown extract resources from the colony's economy. In the first, hierarchical model, the economy is embedded into a moral hazard model of political agency (Ferejohn, 1986). While necessary for military security and economic administration, the colonial governor has incentives to extract rents from settlers that, in turn, decrease the economic investment of settlers and the Crown's revenues. In the separation of powers model, the Crown chooses to let the settlers determine a budget for the government. This constrains the governor and ensures high investments from the settlers, which, in turn, increase the Crown's revenue. Gailmard then shows that separation of powers is optimal for the Crown when returns to investment are moderate. Specifically, he establishes the conditions for the environment (parameters of the model) under which the Crown's choice of separation-of-power institutions (strategy) is the equilibrium (point prediction).

Gailmard's analytic narrative approach offers a new way of looking at the evolution of an important, widespread institution. Interestingly, the model suggests that separation of powers was neither invoked to control the crown (by an assembly), nor invented by the crown to tie its own hands. Rather it was designed to empower settlers to restrain the governor. The formal modeling approach also helps address the reasons that this distinct form of separation of powers was unique to North America? An investigation of the model parameters suggests that, unlike colonies that provided opportunities for natural resource extraction, economic growth in the North American colonies required settler investment in agriculture. In an even broader perspective, the approach offers insights on American institutionalism but may also be taken up by research in a range of other, colonial or authoritarian, contexts of institutional choice.

It should also be clear that, with analytic narratives, theoretical and empirical work are not delegated to distinct steps in the research process, but rather developed more in a dialogue.

Comparative Statics

Comparative statics ask how an equilibrium quantity of interest changes as some exogenous features change (Silberberg & Suen, 2000). As we have argued above (Figure 1), these quantities of interest are outcomes, f, in a broad sense: features that are associated with the real world and that we argue or predict to occur or vary conditional on parameters θ such as actions, strategies or some policy (or outcome). For example, in the entry game above, we could be interested in how the probability that both firms enter the market will change as profits increase. The answer is straightforward. In the mixed strategy equilibrium, the probability that both firms enter is $p_1 p_2 = 4(1 - \frac{(c_1+c_2)}{\Pi} + \frac{c_1 c_2}{\Pi^2})$. Then the positive partial derivative of $p_1 p_2$ with respect to Π,

$$\frac{\partial p_1 p_2}{\partial \Pi} = 4\left(\frac{c_1 + c_2}{\Pi^2} - 2\frac{c_1 c_2}{\Pi^3}\right) = 4\left(\frac{c_1(\Pi - c_2) + c_2(\Pi - c_1)}{\Pi^3}\right),$$

tells us that, when profits increase, it is more likely that both firms enter the market. Less straightforward though, it also holds that sensitivity to profits varies with entry costs. To be precise

$$\frac{\partial^2 p_1 p_2}{\partial \Pi \partial c_1} = \frac{4(\Pi - 2c_2)}{\Pi^3} < 0,$$

as 1's costs of entry increase, the incentive provided by higher profits weakens. The simple example demonstrates the reason that comparative statics is attractive. First, it is the tool for the 'comparative analysis' of formal models, whether used in case study, comparative, experimental or statistical method design. Second, it is often difficult to exactly measure model parameters as they are unobserved or latent (see example in section 3). As long as we know that two cases differ with respect to some parameter, comparative statics can tell us whether and how equilibria or outcomes differ.

Comparative statics are often tricky because, when one parameter, such as profit or cost, changes, the resulting change in the strategy of one actor also leads to changes in the strategies of other actors. Solving these models often involves using systems of differentiated equations.

A standard technique to derive comparative statics is to assume specific functional forms about the model primitives, such as that utility is linear in profits and costs (as above), or concave in profits but convex in costs. The reason is that standard techniques to find equilibria and to study their comparative statics usually require continuity of best responses, compactness of strategy spaces, differentiability of utility functions and interior solutions, among others.

Montone comparative statics (Milgrom & Shannon, 1994) explores properties of games that make such strong assumptions or explicit functional forms superfluous. As Ashworth & Bueno de Mesquita (2006, 214-5) note, monotone comparative statics greatly facilitates the empirical evaluation of theoretical models. On the theory side, not having to assume specific functional forms makes the model robust against many sorts of misspecification, and brings the deep structure of the model to light. On the empirical side, replacing compact strategy spaces (the standard assumption) with partially ordered spaces allows for the generation and evaluation of predictions that are based on just ordinal information. More recently, monotone comparative statics techniques have been extended to aggregate games (Acemoglu & Jensen, 2013) and distributional comparative statics (Jensen, 2017).

A Full Set of Empirical Implications: Snyder & Stromberg (2010)

Snyder & Stromberg (2010) is an excellent example for how comparative statics can be used. They investigate the effect of political newspaper coverage on policies at the constituency level. Their argument takes the form of a chain of relationship predictions between variables that ultimately link newspaper coverage and policies: increased newspaper coverage increases citizens' level of political information, which strengthens the monitoring of politicians who, in turn, work harder for their constituents, finally resulting in better policies. The authors could have chosen to collect data on congruence and policies to test their key hypothesis and stop there. They didn't. Instead, the authors decided to provide empirical evidence for each step in the causal chain, achieving a close connection between theoretical argument and empirics.

To identify the effect of newspaper coverage on policies, they exploit exogenous variation in the congruence between congressional districts and newspapers. The authors expect that the larger the overlap between these districts, the higher political coverage of local politicians because there exists a higher readership share in the district. As a first empirical step, it is shown that newspapers coverage is indeed increasing in the readership share of a district. The next step in the mechanism postulates that the more extensively newspapers cover local politicians, the better informed the citizens. Using data from the American National Election Study over twenty years, the authors demonstrate that respondents in more congruent districts are more likely to receive their news from newspapers or magazines. Respondents living in more congruent districts are also informed about their local representatives: they are more likely to recall the name of at least one representative and are more willing to place them on ideological scales. Following the causal chain argument, the authors continue to show that congressmen from more congruent districts vote more often against the party line and are more likely to stand witness before congressional hearings and to serve on constituency-oriented committees. Finally, as a last step, the authors provide evidence that more congruent districts receive higher federal expenditures.

While Snyder & Stromberg do not present a formal model for their argument, their work stands out as a demonstration of how a series of comparative statics can illuminate a theoretical mechanism.

Structural Estimation

Structural estimation turns the logic of the previous two research designs – point predictions and comparative statics – on its head. Instead of using data to scrutinize a theoretical model by evaluating point or relationship predictions, structural approaches assume that the theoretical model is a good approximation of the real world and use data to obtain estimates of model parameters. We see at least two main reasons to derive and estimate structural econometric models.

First, so-called deep model parameters are frequently interesting and useful. For example, political scientists are often interested in the policy preferences of political actors, such as legislators or Supreme Court Justices. However, these policy ideal points are generally not directly observable and have to be estimated. Ideal point models, such as Nominate (cp. Poole, 2005), posit a behavioral model (including assumptions about the dimensionality of the policy space) that links deep parameters (here: ideal points) to observables (here: roll call votes). Then, the model can be estimated using roll call data, and the actors' ideal points can be recovered and find use in many contexts (for example, the study of polarization in U.S. politics). Other examples of interesting deep parameters are risk aversion, discount factors, marginal cost and the value of information (see first example below, Iaryczower & Shum, 2012).

Second, structural models can be used to assess the effect of hypothetical policy interventions by conducting simulations. Having described and estimated a structural model, a researcher can treat the estimated parameters as fixed and makes predictions about equilibrium outcomes as the environment changes. For example, a researcher may be interested in the effect of a reform that decreases the minimum required sentence on pre-trial bargaining in criminal cases (see second example below, Silveira, 2017). In many cases, it is desirable to obtain estimates of effects of policy interventions *ex ante* because small-scale field experiments are infeasible or unethical. In other cases, structural models can help to extrapolate results from a specific (field) experimental setting to other contexts (Wolpin, 2013).

To demonstrate a structural estimation approach, consider again the entry game example (also discussed in Ellickson & Misra, 2011). Imagine we have collected data of local markets including information on entry decisions of each player as well as firm- and market-specific characteristics (e.g. population, distance to nearest distribution center) that may affect each firm's payoffs. Let X_k denote characteristics of market k; Z_{ik} denote characteristics of firm i in market k; and y_{ik} the entry decision of player i in market k. Then we can express the payoff of firm i in market k as

$$\pi_{ik} = \alpha_i' X_k + \beta' Z_{ik} + \delta_i y_{-ik} + \epsilon_{ik}$$

where ϵ_{ik} is a component of a firm's payoff that is unobservable to the researcher. In equilibrium, each firm i will choose to enter the market k if $\pi_{ik} > 0$ and stay out otherwise. Note that the payoff depends not only on firm and market-specific characteristics, but also on the entry choice of the other player, y_{-ik}.

If we further assume that firms (but not the researcher) observe the ϵs and that firms make choices simultaneously, the Nash equilibrium can be characterized by the following system of inequalities:

$$\begin{aligned} y_{1k} &= \mathbb{I}[\alpha_1' X_k + \beta_1' Z_{1k} + \delta_1 y_{2k} + \epsilon_{1k} \geq 0] \\ y_{2k} &= \mathbb{I}[\alpha_2' X_k + \beta_2' Z_{2k} + \delta_2 y_{1k} + \epsilon_{2k} \geq 0]. \end{aligned}$$

Note that these outcome equations constitute a binary simultaneous equation system, an interdependent structure that creates challenges for estimation and identification.

Estimation can usually not be achieved with econometric techniques that are implemented in standard econometric software. The reason is that, in contrast to single agent decision-making problems, the payoff of each player depends on the action of the other player (note the $\delta_i y_{-ik}$ term in the inequalities above). Any estimation approach must therefore simultaneously estimate both equations. In general, structural approaches often require that researchers tailor estimators precisely to the empirical case.

An identification problem arising in many structural estimation approaches, in general, and in this setting, in particular, is that the estimation procedure needs to deal with equilibrium multiplicity. As shown earlier, the entry game has two pure strategy Nash equilibria if the market is only large enough to make one firm entry profitable but not two. Since, given a set of parameter values, two different outcomes are possible (either firm 1 enters the market and firm 2 does not, or vice versa), parameter estimation is not identified. To deal with this problem, the researcher could either specify an equilibrium selection rule (Tamer, 2003); aggregate to model predictions that are not affected by multiplicity (e.g. number of entrants, see Bresnahan & Reiss, 1990); make different model assumptions that yield unique predictions (e.g. sequential moves, see Berry, 1992); or adopt a bounds approach (Ciliberto & Tamer, 2009). An alternative would be to assume that firms do not perfectly observe each other's payoffs, a situation which would result in a game with incomplete information (Rust, 1994; Seim, 2006).

Estimating Deep Parameters: Iaryczower and Shum (2012)

Iaryczower & Shum (2012) estimate the value of information, i.e. how much information to the contrary is necessary to overcome ideological predispositions of Justices in the U.S. Supreme Court. This value is not directly observable in the real world, but a quantity that can only be interpreted in the context of the theoretical model. Iaryczower & Shum set up the interaction between justices as a Bayesian game in which biased justices receive a private, noisy signal about the case, update their beliefs and vote in favor or against the defendant. Voting can be either sincere or strategic (as in Feddersen & Pesendorfer, 1996) .

The key issue to recover the value of information is to estimate two deep parameters: judge-specific preferences (bias) and information (signal precision). Iaryczower & Shum propose the following two-step estimation: first, the authors estimate, for each judge, choice probabilities for each state of the world (guilty or innocent defendant) as a function of judge and case-specific observables via maximum likelihood. In a second step, bias and precision parameters can be recovered given estimated choice probabilities. Finally, given bias and precision parameters, one can calculate the probability that a justice votes differently than he or she would have voted without case-specific information, which the authors interpret as a measure of the value of information in the court. Tracing this measure over time, the authors find that the value of information has decreased over the last 25 years, suggesting an increasing politicization of the Supreme Court.

Iaryczower & Shum's approach offers an alternative to purely ideological characterizations of Supreme Court justices' behavior by specifically modeling and estimating how ideology interacts with the information available to the justices. Doing so allows new insights in the relative weights that justices put on their preexisting ideological leanings versus the information of the case.

Simulating Counterfactual Scenarios: Silveira (2017)

Silveira (2017) investigates the effects of several hypothetical policy interventions, namely potential sentencing reforms, on the outcomes of criminal cases. Litigation is modeled as a two-stage game. The first stage consists of pre-trial bargaining with asymmetric information between a prosecutor and the defendant, reflecting the fact that the vast majority of criminal cases in the U.S. are resolved by plea bargaining. Defendants, because they know the full extent of their culpability, have more precise information about potential trial outcomes than the prosecutor. The prosecutor offers the defendant a sentence to settle the case as a take-it-or-leave-it offer. If the defendant accepts the offer, the game ends. Otherwise, the case is decided at trial at which the defendant's private information is revealed with some probability. In equilibrium, the prosecutor's offer is a function of the anticipated trial sentence, the trial costs and the perceived odds of winning. The defendant accepts the offer if he or she is comparably pessimistic about his chances in court.

To estimate the model, the author proceeds in two steps. First, he nonparametrically estimates the prosecutor's equilibrium settlement offer function using information on the prosecutor's settlement offers when plea bargaining was successful, on sentences assigned at trial when plea bargaining failed and on the probability that plea bargaining was successful. Having identified the prosecutor's optimal offer function, the model primitives (i.e. the distribution of potential sentences at trial, the probability of conviction and the trial costs for defendant and prosecution) can be recovered. The author uses these estimates to conduct counterfactual simulations to explore the effects of lower mandatory minimum sentences and a general ban of plea bargaining. In the first case, a decrease of the minimum mandatory sentence leads to a general decrease of incarceration time due to the fact that prosecutors offer lower settlements during plea bargaining. However, it also leads to an increase in conviction rates because shorter potential trial sentences increase the probability of a successful plea bargain. In the case of a complete elimination of plea bargaining, all cases would go to trial. The model suggests that conviction rates would decrease substantially, however, defendants would face longer sentences. Thus, the average defendant would be worse off if plea bargaining were to be banned.

Silveira's analysis yields insights in potential sentencing and litigation reforms that would be challenging to assess otherwise *ex ante*. Furthermore, experimental research that could provide *ex post* evidence about the effects could be difficult to implement and is potentially unethical.

Practical Advice

The development of internal specialization in political science is not a new phenomenon. More than 60 years ago, the APSA Committee on Standards of Instructions noted that 'the political scientist increasingly finds the body of political knowledge so great, and the tools of study so exacting, that he must specialize if he is to master and to communicate his subject matter' (APSA

Committee on Standards of Instruction, 1962, 417). Nor has there been a shortage of ideas for methodological unification, with the accomplishments of the Cowles Commission in economics dating all the way back to the 1930s. The EITM approach is a genuine political science approach to break down the isolation of formal and empirical modeling, tailored to the specific questions and methodological problems in political science research.

It would be the best of all possible worlds if researchers would be highly skilled in both theoretical and empirical analysis paired with deep knowledge of the substantive field of their interest. However, both becoming a skillful formal modeler or data analyst requires an enormous amount of training and the time constraints of PhD education usually prevent students from becoming experts in both fields. Most PhD students become one of two types: the formal modeler who would like to explore the data or the empiricist who would like to add more depth to her empirical analysis.

Also, career-wise, it can be more productive to persuade one audience than to try to appeal to two at the same time. This applies to both faculty members and reviewers. As editor and reviewers we saw many excellent papers struggling in the review process because either the theory section did not persuade the theorists or the empirics did not convince the empiricists. Given space constraints in journals, authors often end up cutting the theoretical or the empirical part of their paper, or relegating large parts to the appendix in the revisions.

For both types of students, we want to offer some brief (hopefully helpful) advice from our experience as researchers, teachers, editor and student.

Not every model is for empirical 'testing.' (Nor does the EITM approach suggest so.) Models serve different purposes. They can be foundational, organizational, explanatory or predictive (Clarke & Primo, 2012). Being apt for the derivation of testable implications is one purpose, but testability in the hypothetico-deductive methodology is not the only one.

Ultimately, models are just tools to help us understand the world.

Not every empirical analysis needs a formal model. For some research questions, the mechanisms behind the data is not of primary interests, at least, not in the first step of research. Whether tobacco or gun control programs have causal effects is a relevant question to answer, even if we do not know how policies change individual perceptions, attitudes and behavior. At the same time, without proper theory of what is going on, simply relying on identification techniques like instrumentable variables or regression discontinuity designs can lead to the misinterpretation of the causal effect. As Eggers (2017) shows, there is no such thing as *the* incumbency effect estimated from regression discontinuity.

Collaboration. Where true versatility is rare, collaboration seems to be a logical consequence. However, co-authoring does not mean everyone does her own thing. At a minimum, students must learn each other's mindsets and languages, which is different than just taking a game theory and an advanced statistics class. For example, when theorists write formal models, they will often emphasize existence of equilibria, generalizability of functional forms and mathematical elegance over whether their model generates sharp, relevant and empirically falsifiable predictions. It will help formal modelers to understand the interests of empirical researchers are interested in and to write models that are accessible for empirical researchers, provide intuition and the reasons that we should care.

Start simple. Armored with lots of detailed case knowledge, the empiricist begins writing a model that resembles the case she observes in the real world as closely as possible. As a result, the model becomes quickly extensive, hard or impossible to solve and adds little to the understanding of the theoretical argument. To make it solvable, the empiricist will be forced to make some heroic assumptions that draw criticism from theorists and empiricists alike.

The empiricist, we argue, should have started with the core of the theoretical argument. In many cases, the core of the argument can be phrased as a trade-off. Should I protest or stay at home? Should I turn out to vote or not? The starting point, then, is to build the simplest reasonable model that formalizes this trade-off. And make it more complicated, if necessary, from there.

Notes

[1]Others, like Johnson (2014), are even more skeptical, rejecting entirely the idea that models can be subjected to empirical evaluation. For him, models are fables to derive a principle from. By rendering some abstract concept like 'power' or 'deterrence' more concrete, models help to understand what the concept means in the particular circumstance captured by the model (Johnson, 2019, e7).

[2]It is important to emphasize that the (unknown) benchmark is systematic, explainable variation: a model to predict school choice that uses features like parental education, socio-economic status, and personal traits, and that has an R-square of 5 percent, arguably is mediocre. A stock market model with an R-square of 1 percent can earn you a fortune.

[3]Or, if parsimony is a goal, what should be left out.

[4]The term separability goes back to Sono (1945) and Leontief (1947) in the study of production decisions, where separability is the unaffectedness of the ease of substitution between two factors by a third factor.

[5]If unambiguous, we suppress indices for the unit of analysis.

[6]We deliberately assume that in case of equality, the offer is accepted. Such knife-edge cases are practically irrelevant, as long as (some of) the model parameters have continuous support. This is not the case in, for instance, a standard 2-by-2 pure coordination game where players are indifferent between coordinating on standing on the left or the right of an escalator.

[7]One other assumption we must make is a value for σ. As we do not directly observe y^*, the αs and βs (see below), together with σ, are not uniquely identifiable. A standard and inconsequential assumption is that $\sigma = 1$.

[8]A weaker assumption that captures the case of linear relationships would be to assume that all partial derivatives of the us with respect to X and Z are positive, and all higher-order cross-partials to be zero. The statistical model would then be $\Pr(y = 1|\pi, X, Z) =$

$\Phi(\pi f(X,Z) - g(X,Z) - C)$ for some appropriate f and g that is amenable to non-parametric estimation.

[9]Here, we think of mixed strategies as literal randomizations over pure strategies, but note that there are alternative interpretations (Harsanyi, 1973; Rosenthal, 1979).

References

Acemoglu, D. & Jensen, M. K. (2013). Aggregate comparative statics. *Games and Economic Behavior*, 81, 27–49.

Aldrich, J. H., Alt, J. E., & Lupia, A. (2008). The EITM approach: Origins and interpretations. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 828–843). Oxford: Oxford University Press.

APSA Committee on Standards of Instruction (1962). Political science as a discipline. *American Political Science Review*, 56(2), 417–421.

Ashworth, S. & Bueno de Mesquita, E. (2006). Monotone comparative statics for models of politics. *American Journal of Political Science*, 50(1), 214–231.

Bates, R. H., Greif, A., Levi, M., & Rosenthal, J.-L. (1998). *Analytic narratives*. Princeton, NJ: Princeton University Press.

Battaglini, M., Morton, R. B., & Palfrey, T. R. (2010). The swing voter's curse in the laboratory. *The Review of Economic Studies*, 77(1), 61–89.

Berry, S. & Reiss, P. (2007). Empirical models of entry and market structure. In M. Armstrong & R. Porter (Eds.), *Handbook of Industrial Organization*, volume 3 (pp. 1845–1886). North-Holland: Elsevier.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica*, 60(4), 889–917.

Blackorby, C., Primont, D., & Russell, R. R. (1998). Separability: A survey. In S. Barberà, P. J. Hammond, & C. Seidl (Eds.), *Handbook of utility theory*, volume 1 (pp. 49–92). Dordrecht: Kluwer.

Bresnahan, T. F. & Reiss, P. C. (1990). Entry in monopoly market. *The Review of Economic Studies*, 57(4), 531–553.

Ciliberto, F. & Tamer, E. (2009). Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6), 1791–1828.

Clarke, K. A. & Primo, D. M. (2007). Modernizing political science: A model-based approach. *Perspectives on Politics*, 5(4), 741–753.

Clarke, K. A. & Primo, D. M. (2012). *A model discipline: Political science and the logic of representations*. Oxford University Press.

Eggers, A. C. (2017). Quality-based explanations of incumbency effects. *The Journal of Politics*, 79(4), 1315–1328.

Ellickson, P. B. & Misra, S. (2011). Estimating discrete games. *University of Rochester, Working Paper*.

Feddersen, T. J. & Pesendorfer, W. (1996). The swing voter's curse. *The American Economic Review*, 86(3), 408–424.

Ferejohn, J. (1986). Incumbent performance and electoral control. *Public Choice*, 50(1/3), 5–25.

Gailmard, S. (2017). Building a new imperial state: The strategic foundations of separation of powers in america. *American Political Science Review*, 111(4), 668–685.

Giere, R. N. (2010). *Explaining science: A cognitive approach*. Chicago: University of Chicago Press.

Harsanyi, J. C. (1973). Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points. *International Journal of Game Theory*, 2(1), 1–23.

Iaryczower, M. & Shum, M. (2012). The value of information in the court: Get it right, keep it tight. *The American Economic Review*, 102(1), 202–237.

Jensen, M. K. (2017). Distributional comparative statics. *The Review of Economic Studies*, 85(1), 581–610.

Johnson, J. (2014). Models among the political theorists. *American Journal of Political Science*, 58(3), 547–560.

Johnson, J. (2019). Formal models in political science: Conceptual, not empirical. *The Journal of Politics*, 81(1), e6–e10.

Keane, M. P. & Wolpin, K. I. (2007). Exploring the usefulness of a nonrandom holdout sample for model validation: Welfare effects on female behavior. *International Economic Review*, 48(4), 1351–1378.

Leontief, W. (1947). Introduction to a theory of the internal structure of functional relationships. *Econometrica (pre-1986)*, 15(4), 361.

Low, H. & Meghir, C. (2017). The use of structural models in econometrics. *Journal of Economic Perspectives*, 31(2), 33–58.

McCubbins, M. D. & Schwartz, T. (1984). Congressional Oversight Overlooked: Police Patrols versus Fire Alarms. *American Journal of Political Science*, 28(1), 165–179.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). New York: Academic Press.

McKelvey, R. D. & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10, 6–38.

Milgrom, P. & Shannon, C. (1994). Monotone comparative statics. *Econometrica*, 62(1), 157–180.

Morton, R. B. (1999). *Methods and models. A guide to the empirical analysis of formal models in political science.* Cambridge: Cambridge University Press.

Morton, R. B. (2003). *Methods and models. A guide to the empirical analysis of formal models in political science.* Cambridge: Cambridge University Press.

Mullainathan, S. & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87–106.

National Science Foundation (2002). *Empirical Implications of Theoretical Models Report.* Arlington, Virginia.

Poole, K. T. (2005). *Spatial Models of Parliamentary Voting.* Cambridge: Cambridge University Press.

Popper, K. R. (1959). *The logic of scientific discovery.* New York: Basic Books.

Rosenthal, R. W. (1979). Sequences of games with varying opponents. *Econometrica*, 47(6), 1353–1366.

Rust, J. (1994). Structural estimation of markov decision processes. In R. F. Engle & D. L. McFadden (Eds.), *Handbook of econometrics*, volume 4 (pp. 3081–3143). North-Holland: Elsevier.

Seim, K. (2006). An empirical model of firm entry with endogenous product-type choices. *The RAND Journal of Economics*, 37(3), 619–640.

Signorino, C. S. (1999). Strategic interaction and the statistical analysis of international conflict. *American Political Science Review*, 93(2), 279–297.

Signorino, C. S. (2003). Structure and uncertainty in discrete choice models. *Political Analysis*, 11(4), 316–344.

Silberberg, E. & Suen, W. C. (2000). *The Structure of Economics: A Mathematical Analysis*. New York: McGraw-Hill, 3 edition.

Silveira, B. S. (2017). Bargaining with asymmetric information: An empirical study of plea negotiations. *Econometrica*, 85(2), 419–452.

Smith, A. (1999). Testing theories of strategic choice: The example of crisis escalation. *American Journal of Political Science*, 43(4), 1254–1283.

Snyder, J. M. J. & Stromberg, D. (2010). Press coverage and political accountability. *Journal of Political Economy*, 118(2), 355–408.

Sono, M. (1945). The effect of price changes on the demand and supply of separable goods (in japanese). *Kokumin Keizai Zasshi*, 74, 1–51.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies*, 70(1), 147–165.

Wolpin, K. I. (2013). *The limits of inference without theory*. Cambridge, MA: MIT Press.