Article

EJIR

European Journal of
International Relations
19(3) 427–457

DOI: 10.1177/1354066113494320
ejt.sagepub.com
SAGE

Leaving theory behind: Why simplistic hypothesis testing is bad for International Relations

John J. Mearsheimer
University of Chicago, USA

Stephen M. Walt
Harvard University, USA

Abstract
Theory creating and hypothesis testing are both critical components of social science, but the former is ultimately more important. Yet, in recent years, International Relations scholars have devoted less effort to creating and refining theories or using theory to guide empirical research. Instead, they increasingly focus on 'simplistic hypothesis testing,' which emphasizes discovering well-verified empirical regularities. Privileging simplistic hypothesis testing is a mistake, however, because insufficient attention to theory leads to misspecified empirical models or misleading measures of key concepts. In addition, the poor quality of much of the data in International Relations makes it less likely that these efforts will produce cumulative knowledge. This shift away from theory and toward simplistic hypothesis testing reflects a long-standing desire to professionalize and expand the International Relations field as well as the short-term career incentives of individual scholars. This tendency is also widening the gap between the ivory tower and the real world, making International Relations scholarship less useful to policymakers and concerned citizens. Unfortunately, this trend is likely to continue unless there is a collective decision to alter prevailing academic incentives.

Keywords
cumulative knowledge, hypothesis testing, methodology, policy-relevance, professional norms, scientific realism, theory

Corresponding author:
John J. Mearsheimer, Political Science Department, University of Chicago, 5828 S. University Avenue, Chicago, IL 60637, USA.
Email: j-mearsheimer@uchicago.edu

Introduction

Theory is the lodestone in the field of International Relations (IR). Its theorists are the field's most famous and prestigious scholars. For example, the *TRIP Survey of International Relations Scholars* published in 2009 found that the three scholars 'whose work has had the greatest influence on the field of IR in the past 20 years' were Robert Keohane, Kenneth Waltz, and Alexander Wendt. All three are major theorists whose reputations rest on ideas they have advanced rather than on their empirical work. Almost all of the other scholars on the list — including Bruce Bueno de Mesquita, Barry Buzan, Martha Finnemore, Samuel Huntington, Robert Jervis, Peter Katzenstein, Stephen Krasner, and Susan Strange — are figures who developed ideas that have shaped the research agenda in IR and in some cases influenced policy debates (Jordan et al., 2009: 43, 45, 47).[1] Several of these individuals have done substantial empirical work to support their theories, but their core theoretical ideas account for their stature.

Moreover, virtually all of the classic IR books are theory-laden works like Hans Morgenthau's *Politics among Nations*, Kenneth Waltz's *Theory of International Politics*, Thomas Schelling's *The Strategy of Conflict*, Hedley Bull's *The Anarchical Society*, Robert Keohane's *After Hegemony*, and Alexander Wendt's *Social Theory of International Politics*, among others. The same is true regarding articles, where the landscape is dominated by well-known pieces like John Ruggie's 1982 article on 'embedded liberalism' in *International Organization*, Michael Doyle's 1983 piece on 'Kant, Liberal Legacies and Foreign Affairs' in *Philosophy and Public Affairs*, and James Fearon's 1995 *International Organization* article on 'Rationalist Explanations for War.'

Finally, a body of grand theories — or what are sometimes called the 'isms' — has long shaped the study of international politics. The most prominent among them are constructivism, liberalism, Marxism, and realism. A recent article by several authors of the Teaching, Research, and International Policy (TRIP) surveys nicely summarizes the influence of these families of theory: 'US graduate seminars are littered with readings that advance and critique the various "isms" in IR theory…. Similarly, introductory IR courses and textbooks for undergraduates are often organized around these paradigms.' They add: 'The view of the field as organized largely by paradigm is replicated in the classroom…. Together, realism and liberalism still comprise more than 40% of introductory IR course content at US universities and colleges today, according to the people who teach those classes' (Maliniak et al., 2011: 441, 444). In short, theory is paramount in the IR world.

Yet, paradoxically, the amount of serious attention IR scholars in the United States pay to theory is declining and seems likely to drop further in the years ahead. Specifically, the field is moving away from developing or carefully employing theories and instead emphasizing what we call simplistic hypothesis testing. Theory usually plays a minor role in this enterprise, with most of the effort devoted to collecting data and testing empirical hypotheses.[2]

This trend is reflected in the TRIP surveys. Although fewer than half of IR scholars primarily employ quantitative methods, 'more articles published in the major journals employ quantitative methods than any other approach.' Indeed, 'the percentage of articles using quantitative methods is vastly disproportional to the actual number of scholars who identify statistical techniques as their primary methodology.' Recent American

Political Science Association (APSA) job postings in IR reveal a strong preference for candidates with methodological expertise and hardly any job postings for theorists. The TRIP survey authors suggest that a 'strong bias' in favor of quantitative methods 'may explain why junior scholars are increasingly trained to use statistics as their primary methodological approach' (Maliniak et al., 2011: 439, 453).

The growing emphasis on methods at the expense of theory is especially pronounced in the subfield of international political economy (IPE). Surveying its history over the past four decades, Benjamin Cohen (2010: 887) notes that 'the character of what gets published in leading journals in the United States … has changed dramatically.' What now fills the pages of those journals is research that makes 'use of the most rigorous and up-to-date statistical methodologies' (also see Oatley, 2011; Weaver et al., 2009). Theoretical debates, which once occupied such a prominent role in the IPE literature, have diminished in importance.

Indeed, some senior IR scholars now rail against the field's grand theories. In his 2010 International Studies Association (ISA) presidential address, for example, David Lake described the 'isms' as 'sects' and 'pathologies' that divert attention away from 'studying things that matter' (Lake, 2011: 471). Thus, it is not surprising that 'the percentage of non-paradigmatic research has steadily increased from 30% in 1980 to 50% in 2006' (Maliniak et al., 2011: 439). Of course, one could advocate for middle-range theories while disparaging grand theories, and indeed Lake does just that. The field is not moving in that direction, however. Nor is it paying more attention to formal or mathematically oriented theories (Bennett et al., 2003: 373–374). Instead, it is paying less attention to theories of all kinds and moving toward simplistic hypothesis testing.

This trend represents the triumph of methods over theory. In recent decades, debates about how to study IR have focused primarily on the merits of qualitative versus quantitative approaches or on the virtues of new methodological techniques. Although not without value, these disputes have diverted attention from the critical role that theory should play in guiding empirical analysis.[3] This focus on methods rather than theory is not the result of a conscious, collective decision by IR scholars, but is instead an unintended consequence of important structural features of the academic world.

The road to ruin

We believe downgrading theory and elevating hypothesis testing is a mistake. This is not to say that generating and testing hypotheses is unimportant. Done properly, it is one of the core activities of social science. Nevertheless, the creation and refinement of theory is the most important activity in this enterprise. This is especially true in IR, due to the inherent complexity and diversity of the international system and the problematic nature of much of the available data. Scholars do not have to invent their own theory, of course, or even refine an existing theory, although these endeavors are highly prized. It is necessary, however, that social scientists have a solid grasp of theory and use it intelligently to guide their research.

Christopher Achen, a prominent methodologist, summarizes what happens when political scientists short-change theory in favor of what he calls 'dreary hypothesis-testing.' 'The present state of the field is troubling,' he writes, 'for all our hard work, we have

yet to give most of our new statistical procedures legitimate theoretical microfoundations, and we have had difficulty with the real task of quantitative work — the discovery of reliable empirical generalizations' (Achen, 2002: 424, 443; also Braumoeller and Sartori, 2004; Schrodt, 2006, 2010; Signorino, 1999).

Theory is invaluable for many reasons. Because the world is infinitely complex, we need mental maps to identify what is important in different domains of human activity. In particular, we need theories to identify the causal mechanisms that explain recurring behavior and how they relate to each other. Finally, well-crafted theories are essential for testing hypotheses properly; seemingly sophisticated tests that are not grounded in theory are likely to produce flawed results.

Our bottom line: de-emphasizing theory and privileging hypothesis testing is not the best way to gain new knowledge about international politics. Both activities are important to scholarly progress, but more attention should be devoted to theory development and hypothesis testing should be tied more closely to theory.

Caveats

This article does not compare the merits of qualitative versus quantitative methods, or argue that qualitative methods are better suited to studying IR. Rather, we argue that theory must play a central role in guiding the research process, regardless of how the theory is tested. We focus primarily on quantitative research because so much of the work in the field now uses this approach. But our arguments apply with equal force to qualitative research and there are numerous examples of qualitative scholarship that devote insufficient attention to theory. Our main concern, in short, is the relationship between theory and empirical work, not the relative merits of quantitative or qualitative approaches.

Nor do we make the case here for any particular IR theory. Although we both work in the realist tradition, we think many kinds of theory — including middle-range theory — can be useful for helping us understand how international politics works. In our view, a diverse theoretical ecosystem is preferable to an intellectual monoculture.

We recognize that the existing body of IR theory contains significant defects, and we are far from nostalgic about some by gone 'Golden Age' where brilliant theorists roamed the earth. There is much work to be done to clarify our existing stock of theories and develop better ones. Nonetheless, we believe progress in the field depends primarily on developing and using theory in sophisticated ways.

We have not read every recent article that tests hypotheses, of course; the current literature is too vast to permit such an exercise. We have read widely, however, and we asked experts who work in the hypothesis-testing tradition to direct us to the best works in this genre. We have also studied assessments of the field that have leveled criticisms similar to ours. The problems we identify are clearly no secret, and some efforts have been made to address them. Contemporary IR research continues to neglect theory, however, and this trend does not bode well for the future of the field.

Regarding epistemology, we focus on so-called positivist approaches to doing IR. Accordingly, we do not discuss critical theory, interpretivism, hermeneutics, and some versions of constructivism. This omission is due in part to space limitations, but also

because our focus is on IR in America, where positivism predominates. As the authors of the TRIP surveys note, 'IR in the United States is overwhelmingly positivist' (Maliniak et al., 2011: 439, 455). There is more epistemological variety outside the United States, especially in Europe, and less emphasis on simplistic hypothesis testing.

In sum: this article is not a *cri de coeur* by two grumpy realists who are opposed to hypothesis testing in general and quantitative analysis in particular. To make our position perfectly clear: we regard hypothesis testing as a core component of good social science. Our point is that this activity must be guided by a sophisticated knowledge of theory and that contemporary IR scholarship is neglecting this requirement.

Our argument is organized as follows. We begin by describing what theories are, why they are essential, and how they should be tested. We also explore the important distinction between scientific realism and instrumentalism, which distinguishes our approach from that of many other positivists. Then we describe simplistic hypothesis testing and the problems that arise from its cursory attention to theory.

Next we consider why IR is moving in this direction despite the significant problems this approach encourages. In this discussion, we explore how the growing emphasis on hypothesis testing makes IR scholarship less relevant for debates in the policy world. Finally, we offer some suggestions on how IR scholars might be encouraged to place more emphasis on theory. It will be difficult to reverse present trends, however, unless the field proves more open to revision than we suspect is the case.

Theory and social science

What is a theory?

Theories are simplified pictures of reality. They explain how the world works in particular domains. In William James's famous phrase, the world around us is one of 'blooming, buzzing confusion': infinitely complex and difficult to comprehend. To make sense of it we need theories, which is to say we need to decide which factors matter most. This step requires us to leave many factors out because they are deemed less important for explaining the phenomena under study. By necessity, theories make the world comprehensible by zeroing in on the most important factors.

Theories, in other words, are like maps. Both aim to simplify a complex reality so we can grasp it better. A highway map of the United States, for example, might include major cities, roads, rivers, mountains, and lakes. But it would leave out many less prominent features, such as individual trees, buildings, or the rivets on the Golden Gate Bridge. Like a theory, a map is an abridged version of reality.

Unlike maps, however, theories provide a causal story. Specifically, a theory says that one or more factors can explain a particular phenomenon. Again, theories are built on simplifying assumptions about which factors matter the most for explaining how the world works. For example, realist theories generally hold that balance-of-power considerations can account for the outbreak of great-power wars and that domestic politics has less explanatory power. Many liberal theories, by contrast, argue the opposite.

The component parts of a theory are sometimes referred to as concepts or variables. A theory says how these key concepts are defined, which involves making assumptions

about the key actors. Theories also identify how independent, intervening, and dependent variables fit together, which enables us to infer testable hypotheses (i.e. how the concepts are expected to covary). Most importantly, a theory *explains* why a particular hypothesis should be true, by identifying the causal mechanisms that produce the expected outcome(s). Those mechanisms — that are often unobservable — are supposed to reflect what is actually happening in the real world.

Theories provide general explanations, which means they apply across space and time. Social science theories are not universal, however; they apply only to particular realms of activity or to specific time periods. The scope of a theory can also vary significantly. Grand theories such as realism or liberalism purport to explain broad patterns of state behavior, while so-called middle-range theories focus on more narrowly defined phenomena like economic sanctions, coercion, and deterrence.

No social science theory explains every relevant case. There will always be a few cases that contradict even our best theories. The reason is simple: a factor omitted from a theory because it normally has little impact occasionally turns out to have significant influence in a particular instance. When this happens, the theory's predictive power is reduced.

Theories vary enormously in their completeness and the care with which they are constructed. In a well-developed theory, the assumptions and key concepts are carefully defined, and clear and rigorous statements stipulate how those concepts relate to each other. The relevant causal mechanisms are well specified, as are the factors that are excluded from the theory. Well-developed theories are falsifiable and offer non-trivial explanations. Finally, such theories yield unambiguous predictions and specify their boundary conditions.

By contrast, casual or poorly developed theories, or what are sometimes called folk theories, are stated in a cursory way. Key concepts are not well defined and the relations between them — to include the causal mechanisms — are loosely specified. The domino theory, which was so influential during the Cold War, is a good example of a folk theory. In our view, much of the hypothesis testing that is done in IR today employs casual or incomplete theories.

Our conception of theory applies with equal force to formal theories, which employ the language of mathematics, and non-formal theories, which use ordinary language. Theories are ultimately acts of imagination and the language in which they are expressed — be it mathematical notation or words — matters less than whether the theory offers important insights into a particular realm of IR. The key criterion is whether the theory has explanatory power, not whether it is formal or non-formal.

On epistemology: Scientific realism versus instrumentalism

To make our views on theory crystal clear, some brief words about epistemology are in order. As some readers have probably recognized, our perspective is that of scientific realism.[4] Theories, for us, comprise statements that accurately reflect how the world operates. They involve entities and processes that exist in the real world. Accordingly, the assumptions that underpin the theory must accurately reflect — or at least reasonably approximate — particular aspects of political life. Assumptions, we believe, can be

shown to be right or wrong and theories should rest on realistic assumptions. They are not 'useful fictions' that help generate interesting theories, as some social scientists claim. For scientific realists, a rational actor assumption makes sense only if the relevant agents in the real world behave strategically. Otherwise, the resulting theory will not have much explanatory power.

Furthermore, the causal story that underpins the theory must also reflect reality. In other words, the causal mechanisms that help produce the actual phenomenon being studied must operate in practice the way they are described in the theory. Of course, there will be unobservable as well as observable mechanisms at play in most theories. Just think about the importance of gravity, an unobservable mechanism that is central to our understanding of the universe. Or consider the role that insecurity plays in many international relations theories. We cannot measure insecurity directly, because it is a mental state we cannot observe. But scholars can often detect evidence of its presence in what leaders do and say. Scientific realists believe that those unobservables must accurately reflect reality for the theory to perform well. In short, not only must a theory's predictions be confirmed by empirical observation, but the observed results must also occur *for the right reasons*, i.e. via the causal logics that flow from the theory's realistic microfoundations.

The main alternative epistemology is instrumentalism. It maintains that a theory's assumptions do not have to conform to reality. Indeed, Milton Friedman (1953) famously asserted that the *less* a theory's assumptions reflect reality, the more powerful that theory is likely to be. In this view, assumptions are simply useful fictions that help generate theories. For example, instrumentalists do not care if actors are rational or not, so long as assuming rationality produces theories that generate accurate predictions. In other words, the utility of a theory's assumptions is determined solely by whether its predictions are confirmed.

Instrumentalists dismiss the idea that theories contain causal mechanisms that reflect what is actually happening in the real world. Their perspective is largely driven by the belief that nothing is gained by focusing on unobservable mechanisms, which are often at the center of the causal process (Chakravartty, 2011: 4). For instrumentalists, science is all about measuring observables, which in turn encourages hypothesis testing.

Instrumentalists recognize that theories should contain clearly defined concepts and be logically consistent. They care about a theory's causal logic insofar as they want to tell a coherent story. But they do not believe that the causal process depicted in a theory necessarily reflects reality.[5] As Paul MacDonald (2003: 555) observes, 'instrumentalists are simply treating theories as devices that generate hypotheses,' where the value of the theory is determined solely by whether the hypotheses are confirmed.

We believe scientific realism is the more convincing epistemology. Instrumentalists ask us to believe that a theory can generate accurate predictions even if its assumptions and causal story are at odds with reality. As MacDonald (2003: 554) notes, 'If a theoretical assumption is a fiction, it is unlikely to be empirically useful unless it generates hypotheses that are right for the wrong reasons.' Or as Hilary Putnam famously says, unless it produces a 'miracle' (1975: 73). By definition, theories exclude a vast number of factors and employ simplifying assumptions about the relevant actors. But a good theory must still offer an accurate — albeit abstracted or simplified — portrayal of the

real world. Maps by necessity simplify reality, but a roadmap that placed Chicago east of Boston would not be useful. Theories will produce sound hypotheses and useful explanations only if their components accurately reflect the real world.

How are theories tested?

There are three ways to evaluate a theory. The first is to inspect its logical soundness. Logical consistency is a prized quality in any theory, even though some valuable theories had logical flaws that were resolved over time.[6]

The second method is covariation, which is where hypothesis testing comes in. Given a theory that says A causes B, the objective is to examine the available evidence to determine whether A and B covary. Correlation is not causation, however, which means that it is necessary to show that A is causing B and not the other way around. It is also necessary to show that some omitted factor C is not causing both A and B. To deal with these issues, researchers rely on various techniques of causal inference, which specify how to draw conclusions about cause and effect from the observed data. In essence, causal inference is correlational analysis, using careful research design and appropriate control variables to tease out the independent causal effects of A on B.[7]

The third way to test a theory is process tracing. Here the aim is to determine whether a theory's causal mechanisms are actually operating in the real world in the manner it depicts.[8] In other words, if a theory maintains that A leads to B for a particular reason, then it should be possible to collect evidence to determine whether that is true. For example, some scholars maintain that democracies do not fight each other because they share a commitment to peaceful resolution of disputes; if so, there should be evidence that whenever two democracies were on the brink of war with each other, they refrained from fighting for that reason (Layne, 1994). In essence, process tracing focuses on examining the accuracy of the explanations that underpin a theory's main predictions.

Process tracing is fundamentally different from the first method, which seeks to determine whether a theory is logically consistent. With process tracing, the aim is to examine the empirical performance of the theory's explanatory logic. In that regard, it is similar to hypothesis testing, which is also concerned with assessing empirical performance.

All three methods are valid ways of assessing theories; in fact, they complement each other. In a perfect world, one would employ all of them, but that approach is not always practical. The methods a scholar uses depend on the nature of the puzzle, the availability of relevant evidence, and his or her own comparative advantage.

In contrast to our view, instrumentalists do not believe that process tracing is a useful way to test theories. For them, making sure a theory is logical and testing its predictions are the only valid ways to assess its worth. It is therefore unsurprising that scholars who rely on statistics to evaluate hypotheses often embrace an instrumentalist epistemology, for what matters is simply whether the independent and dependent variables covary as predicted.

As noted above, no social science theory is 100% accurate. But if a theory is tested against a large number of cases and can account for most of them, our confidence in it increases. If a theory makes one false prediction but others hold up well, we still regard it as useful. Also, a weak theory can sometimes become more useful because conditions

in the real world change. For example, the theory that economic interdependence discourages war may be more valid today than it was in the past because globalization has made it more costly for major powers to fight each other (Brooks, 2007).

Finally, how we think about any theory is ultimately a function of how it compares with its competitors. If we know a theory is flawed but do not have a better one, it makes sense to stick with it despite its defects, because we cannot function without some sort of theory to guide us. A weak theory is better than no theory at all, and flawed theories often provide the point of departure for devising new and better ones.[9]

The virtues of theory

Theory is important for many reasons. First, theories provide overarching frameworks — the 'big picture' — of what is happening in myriad realms of activity. There is simply no way to understand an infinitely complex world just by collecting facts. Carl von Clausewitz (1976: 145, 577–578) saw this clearly: 'Anyone who thought it necessary or even useful to begin the education of a future general with a knowledge of all the details has always been scoffed at as a ridiculous pedant.' He goes on to say, 'No activity of the human mind is possible without a certain stock of ideas.' In other words, we need theories.

Theories, in short, provide economical explanations for a wide array of phenomena. They help us interpret what we observe and tie different hypotheses together, making them more than just a piecemeal collection of findings. This is why economists group theories into schools of thought such as Keynesianism, monetarism, rational expectations, behavioral economics, etc. IR scholars array their theories as 'isms' for much the same reason.

Although theory is necessary in every realm of life, the more complicated and diverse the realm, the more dependent we are on mental maps to help us navigate the terrain. IR should place a high value on theory, therefore, because it seeks to make sense of an especially large and complex universe. As David Lake (2011: 467) notes, 'International studies deals with the largest and most complicated social system possible.' This complexity, he points out, accounts in part for 'the diverse range of research traditions' in the field. Moreover, IR scholars cannot assume that findings obtained in one context will apply in a different one, unless they can invoke a theory that explains why seemingly diverse contexts are sufficiently similar. For these reasons, IR is more dependent on theory than other fields in political science or the social sciences more generally.

Second, powerful theories can revolutionize our thinking. They transform our understanding of important issues and explain puzzles that made little sense before the theory was available. Consider Charles Darwin's impact on how people thought about the origins of the human species and countless other phenomena. Before Darwin published his seminal work on evolution, most people believed that God played the key role in creating humankind. Darwin's theory undermined that view and caused many people to change their thinking about God, religion, and the nature of life itself.

On a lesser scale, consider the phenomenon of free-riding, which plagues many types of collective action. This seemingly puzzling form of behavior was clarified when Mancur Olson (1965) and others explained why free-riding is perfectly rational in many

circumstances. This new knowledge also alters subsequent behavior, for once people understand Olson's logic, their incentive to free-ride increases. A handful of separate and well-verified hypotheses would have had far less impact than a simple and powerful theory like Darwin's or Olson's.

Third, theory enables prediction, which is essential for the conduct of our daily lives, for policymaking, and for advancing social science. Each of us is constantly making decisions with future consequences and trying to determine the best strategy for achieving desired goals. Simply put, we are trying to predict the future. But because many aspects of the future are unknown, we must rely on theories to predict what is likely to happen if we choose one strategy over another.

Fourth, as should be clear from the previous discussion, theory is essential for diagnosing policy problems and making policy decisions. Government officials often claim that theory is an academic concern and irrelevant for policymaking, but this view is mistaken. In fact, policymakers have to rely on theory because they are trying to shape the future, which means that they are making decisions they hope will lead to some desired outcome. In short, they are interested in cause and effect, which is what theory is all about. Policymakers cannot make decisions without at least some vague theory to tell them what results to expect. As Robert Dahl notes: 'To be concerned with policy is to focus on the attempt to produce intended effects. Hence policy-thinking is and must be causality-thinking.'[10]

Fifth, theory is crucial for effective policy evaluation (Chen, 1990). A good theory identifies indicators we can use to determine whether a particular initiative is working, because criteria for evaluation are embedded within it. For example, if one's theory of counterinsurgency suggests that the key to victory is killing large numbers of insurgents, body counts are an obvious benchmark for assessing progress. But if one's theory of victory identifies winning hearts and minds as the key to success, then reliable public opinion polls would be a better indicator. In short, effective policy evaluation depends on good theory.

Sixth, our stock of theories informs retrodiction: theory enables us to look at the past in different ways and better understand our history (Trachtenberg, 2006: ch. 2). For example, the democratic peace hypothesis was barely recognized before the early 1980s, but scholars have subsequently used it to account for periods of peace reaching far back into the past (Doyle, 1983; Weart, 1998). Similarly, the 'cult of the offensive' interpretation of the origins of World War I (Lynn-Jones, 1995; Van Evera, 1984) did not exist before the creation of offense-defense theory in the mid-1970s. Of course, we can also test a new theory by asking what the historical record should show if it is correct. Lastly, new theories by definition provide alternative ways of explaining past events, and thus provide tools for critiquing existing historical accounts.

Seventh, theory is especially helpful when facts are sparse. In the absence of reliable information, we have little choice but to rely on theory to guide our analysis. As Jack Snyder (1984/1985) noted during the Cold War, the dearth of reliable facts about the Soviet Union made it necessary to rely on theory to understand what was going on inside that closed society. There is always the danger, however, that one might apply a familiar theory to a situation for which it is not applicable. Yet, when reliable information is at a premium, we are forced to rely more heavily on theory.

Theory can be particularly valuable for understanding novel situations, where we have few historical precedents to guide our thinking. For example, the invention of nuclear weapons in 1945 created a new set of strategic problems that led to the invention of deterrence theory and other related ideas (Kaplan, 1983: ch. 6; Wohlstetter, 1959). Similarly, novel environmental challenges helped inspire Elinor Ostrom's Nobel Prize-winning work on managing natural resources more effectively (Ostrom, 1990). Lastly, the advent of unipolarity requires us to devise new theories to explain how this new configuration of power will affect world politics (Ikenberry et al., 2011; Monteiro, 2011/2012; Wohlforth, 1999).

Eighth, as discussed at greater length below, theory is critical for conducting valid empirical tests. Hypothesis testing depends on having a well-developed theory; otherwise, any tests we perform are likely to be of limited value. In particular, our stock of theories can suggest causal factors that scholars might not have recognized and thus omitted from their analysis. Furthermore, theories are essential for defining key concepts, operationalizing them, and constructing suitable data sets. One must have a clear understanding of the theory being tested in order to know whether the things being measured or counted accurately reflect the concepts of interest.[11]

In sum, social science consists of developing and testing theory. Both activities are essential to the enterprise. There are two possible dangers, therefore: (1) theorizing that pays too little attention to testing; and (2) empirical tests that pay too little attention to theory. Because any discipline must perform both activities, the key issue is finding the optimal balance between them. As we will now show, the balance in IR has shifted away from theory and toward simplistic hypothesis testing, to the detriment of the field.

What is simplistic hypothesis testing?

At the risk of caricature, simplistic hypothesis testing begins by choosing a particular phenomenon (the dependent variable), which is often a familiar topic like war, alliance behavior, international cooperation, human rights performance, etc. The next step is to identify one or more independent variables that might account for significant variation in the dependent variable. These independent variables can be identified from the existing literature or by inventing a new hypothesis. Each of these hypotheses thus highlights a different possible cause of the phenomenon under study.

The researcher(s) then selects data sets containing measures of the independent and dependent variables, along with whatever control variables are thought to be important for making valid causal inferences. If appropriate data sets do not exist, new ones must be compiled. Finally, the hypotheses are tested against each other, usually with some type of regression model, using various statistical techniques to deal with endogeneity, collinearity, omitted variables, or other sources of bias.

The ultimate aim of this approach is measuring the covariation between the different independent variables and the dependent variable, to determine which independent variables have the greatest causal impact.[12] Large-*N* quantitative analysis is usually the preferred approach, based on the belief that it is the most reliable way of measuring causal influence (King et al., 1994). The desired result is one or more well-verified hypotheses, which become part of a growing body of knowledge about international behavior.

What role does theory play?

For the most part, contemporary hypothesis testers are not engaged in pure induction, mindlessly churning data in search of interesting correlations. Nonetheless, theory plays a modest role in much of this work. Although the hypotheses being tested are sometimes drawn from the existing literature, relatively little attention is paid to explaining how or why a particular independent variable might cause the dependent variable. In other words, little intellectual effort is devoted to applying existing theory carefully; i.e. to identifying the microfoundations and causal logics that underpin the different hypotheses. Nor is much effort devoted to determining how different hypotheses relate to one another or to refining the theory itself.

Instead, the emphasis is on testing the hypotheses themselves. Once a scholar can offer a plausible story for why A might have some effect on B, the next step is to collect data and see if a statistically significant relationship can be found. Scholarship proceeds on the assumption that truth lies in the data, and what matters most is empirical verification. As James Johnson (2010: 282) observes, supporters of this approach 'have reinforced a nearly exclusive, but unjustifiable, focus on empirical performance as the chief, perhaps exclusive, criterion of assessment in social and political inquiry.'

It is worth noting that this approach leads toward de facto instrumentalism. Some hypothesis testers acknowledge the importance of causal mechanisms, but their approach does not seek to specify the mechanisms linking independent and dependent variables and devotes virtually no attention to exploring them directly. Their focus, to repeat, is on measuring covariation. Figuring out *why* an observed association obtains — which is the purpose of theory — gets left behind.

To reiterate: theory plays a background role in contemporary hypothesis testing, in the sense that the hypotheses are often loosely based on prior theoretical work and usually have a certain a priori plausibility. But the emphasis is on testing rival hypotheses with the latest statistical techniques. The balance between theory creation and refinement, on the one hand, and empirical verification, on the other, heavily favors the latter. Nor does theory play a major role in guiding the hypothesis-testing process.

What problems arise from inadequate attention to theory?

Privileging hypothesis testing might make sense if it produced lots of useful knowledge about international relations. This does not appear to be the case, however, even though the number of scholars and publications using this approach has increased significantly. As Achen (2002: 424) notes in a broad critique of methodological practice in political science, 'Even at the most quantitative end of the profession, much contemporary empirical work has little long-term scientific value.' Or, as Beck et al. (2000: 21) point out, 'Despite immense data collections, prestigious journals, and sophisticated analyses, empirical findings in the quantitative literature on international conflict are frequently unsatisfying …. Instead of uncovering new, durable, systematic patterns … students of international conflict are left wrestling with their data to eke out something they can label a finding.' The lack of progress is unsurprising because simplistic hypothesis testing is inherently flawed.

Misspecified models. Models used to test hypotheses are statistical representations of some proposed theory. Accordingly, even a sophisticated hypothesis test will not tell us much if the model does not conform to the relevant theory. In order to conduct valid tests, therefore, we need to understand how the variables in the theory fit together and the hypothesis tests must be designed with the theory's assumptions and structure in mind.

Consider the issue of omitted variables. If an important variable is omitted from a regression model, the other coefficients in the model will be biased. This problem is commonly treated as a methodological issue, but it is actually a theoretical matter. Specifically, to argue that a key variable has been omitted is another way of saying that the underlying theory on which the hypothesis tests are based is incomplete. Like all forms of specification error, the problem is that the statistical model being used to test the hypothesis does not conform to the actual causal relations among the key variables. In such circumstances, large regression coefficients and small standard errors are no guarantee of validity.[13]

The same principle applies to the familiar issue of selection bias. This problem is also commonly treated as a methodological issue, but it occurs because some underlying causal mechanism is affecting the observed data in ways that have not been taken into account by the researcher, thereby biasing estimates of causal impact.

To see this clearly, consider James Fearon's critique of Paul Huth and Bruce Russett's analyses of extended deterrence.[14] Huth and Russett test a number of hypotheses about the factors that make deterrence more effective, focusing on the balance of power and the balance of interests. Like much of the published work in the hypothesis-testing tradition, their results vary depending on the specific model being estimated. For example, in some of their models the impact of nuclear weapons is not statistically significant; in others, possessing nuclear weapons has a positive effect. Huth and Russett find that a favorable balance of forces makes deterrent success more likely, while Huth's more recent work found that the balance of interests did not have much effect on deterrent success (Huth, 1988).

Fearon uses a simple bargaining model to show how states take balances of power and interests into account before entering a crisis, and proceed only when they are reasonably confident of success. In other words, states select themselves into crises, thereby creating the historical record that is being used to test different hypotheses. These selection effects must be taken into account when estimating the impact of these factors on the success or failure of deterrence.

Fearon uses this insight to reinterpret Huth and Russett's data and gets different and more consistent results. The point is that Fearon's underlying theory — his picture of how states interact and how the different elements of deterrence are connected — differs from the theory employed by Huth and Russett. It is this theoretical revision that leads to more convincing empirical findings. As Fearon notes: 'both the construction of data sets and the interpretation of empirical findings tend to be strongly shaped by the implicit or explicit theoretical apparatus employed by the analyst' (1994: 266).

Even when selection bias is not an issue and we have identified the relevant independent variables, we still need theory to tell us how they are related. To take a simple example, if X causes Y via an intervening variable Z, and we insert Z into the regression equation as a control variable, the estimated causal relationship between X and Y will

decrease or disappear. This might lead us to erroneously conclude that X had no effect on Y. Indeed, simply inserting control variables into a statistical model can be problematic if it is done because one suspects that they have some impact on the dependent variable but there is no concrete theoretical basis for this belief. Without good theory, in short, we cannot construct good models or interpret statistical findings correctly.[15]

Moreover, understanding how the variables fit together is essential for selecting appropriate statistical procedures. In other words, you need to know a lot about the underlying theory to know what kind of statistical model to use. Yet as Braumoeller and Sartori (2004: 133, 144–145) point out, many IR scholars do not pay much attention to this issue. In their words, 'Empirical researchers often spend too much effort calculating correlations with little or no attention to theory … [and] often impose a statistical model on the theory instead of crafting a model to test the theory.' In particular, the linear regression model that is commonly used to test hypotheses yields incorrect results when the relationship among the key variables is non-linear, conjunctural, or reciprocal.

For example, if the relationship between democratization and war is curvilinear (Mansfield and Snyder, 2007), testing this hypothesis with a linear model is likely to yield biased results. As Philip Schrodt (2006: 337) warns, 'for many data sets commonly encountered in political research, linear models are not just bad, they are really, really bad.'

Or as Achen (2005: 336) observes:

> Garbage-can lists of variables entered linearly into regression, probit, logit and other statistical models have no explanatory power without further argument. Just dropping variables into SPSS, STATA, S or R programs accomplishes nothing, no matter how high-powered or novel the estimators. In the absence of careful supporting argument, the results belong in the statistical rubbish bin.

Misleading measures. Valid hypothesis tests depend on having measures that correspond to the underlying concepts being studied. This requires careful attention to theory, to ensure that key concepts are defined precisely and the indicators used to measure them reflect the concepts as well as the causal relations depicted in the theory.

Unfortunately, contemporary IR scholarship faces challenging measurement issues, due in part to inadequate attention to theory. For example, Alexander Downes and Todd Sechser (2012) show that hypothesis tests that appear to confirm the impact of audience costs had measured several key concepts in ways that did not correspond to the logic of the theory. According to audience costs theory, democratic states in a crisis make more credible threats than authoritarian regimes do, because democratic leaders know they will pay a political price if they back down in public. This concern makes them less likely to bluff, so any threats they make should be taken more seriously and be more effective than threats made by autocrats.

Given the theory's logic, testing it properly requires comparing the effectiveness of explicit public threats issued by key officials in democratic and authoritarian regimes. Measures of the dependent variable must also identify the outcome of each confrontation and whether the target(s) of a given threat complied with it or not. Unfortunately, the data sets previously used to test the theory — the well-known Militarized Interstate Dispute (MID) and International Crisis Behavior (ICB) data sets — do not meet either criterion.

In particular, they: (1) include many crises where no explicit threats were made; (2) include threatening actions unauthorized by national leaders; and (3) code crisis outcomes in ways that do not identify whether the threats were successful or not. When more appropriate data are employed, audience costs do not appear to give democratic leaders any bargaining advantage.

Dan Reiter and Allan Stam's (2002) *Democracies at War* offers another example of a sophisticated study that nonetheless contains questionable measures of key concepts. They argue that democracies perform better in war in part because they have a 'liberal political culture' that encourages individualism, which in turn produces soldiers who exhibit greater initiative in battle. Their empirical analysis appears to support this claim, but the measures they employ to test this idea do not capture the theory's core concepts.

As Risa Brooks (2003) points out, Reiter and Stam measure 'liberal political culture' using regime-type scores from the POLITY III data set. Yet this data set does not contain any direct measure of political culture, let alone liberalism. Rather, it codes a state's level of democracy by measuring electoral competitiveness and other institutional features. Because states can be formally democratic but not liberal, a high score on the POLITY III index is at best loosely related to the concept — 'liberal political culture' — that supposedly determines military performance. To make matters worse, Reiter and Stam measure 'initiative' by using a data set that appears to code which commander(s) launched the first attack in a given battle. This indicator, however, would not measure the initiative displayed by small units or individual soldiers, which is the variable on which their argument depends.

To be fair, these measurement issues are partly due to the conceptual complexity of international politics itself. IR scholars do not have straightforward ways to measure many key concepts or even general agreement on how these concepts should be defined. For example, there is no consensus on how national power should be conceptualized or what the best measure for it might be. Similar problems arise with concepts such as polarity, coercion, or international cooperation. Because rigorous tests using vague concepts will not take us very far, the IR field should place as much value on refining concepts and figuring out how to measure them as it places on hypothesis testing itself. Once again, we see the inescapable need for theory.

Poor data. Privileging hypothesis testing is also unwise given the low quality of much of the data in IR and the importance much of the field assigns to phenomena that are rare or have never occurred. In a perfect world, we would test hypotheses with an abundance of highly reliable data. But in contrast to a field like voting behavior where reliable data are plentiful, data in much of IR are poor. Consider, for example, that contemporary estimates of excess civilian deaths resulting from the 2003 US invasion of Iraq range from under 100,000 to roughly 1.2 million, even though this conflict received enormous attention (Tapp et al., 2008). If the Iraq war is subject to such uncertainty, can we trust the standard IR data sets, especially when dealing with the distant past? In fact, despite a great deal of serious scholarly effort, existing data sets on relative power, terrorism, human rights performance, and a host of other topics are still of questionable reliability.[16]

To make matters worse, much of the raw data that go into standard IR data sets are generated by different agencies in different countries and in many cases are not directly comparable. Even a seemingly straightforward measure such as defense spending cannot be directly compared across countries, because each state includes different items under that heading and calculates the figure differently (Van Evera, 2009). IR scholars are aware of these problems and have worked to address them, but impressive limits to the available data remain.

These data problems can lead to questionable research practices. As discussed above, scholars lacking good data for a key variable may end up using whatever indicators are readily available, even if they do not capture the relevant concepts. Moreover, if scholars follow the frequent admonition to maximize observations, they may include cases for which the data are poor instead of analyzing a smaller number of cases where the data are more reliable.[17]

Lastly, hypothesis testing in IR is constrained when dealing with phenomena where the universe of cases is small or even non-existent, as in the case of social revolution or nuclear war. Standard statistical methods will not work in these situations (Beck et al., 2000), forcing scholars to rely on theory, qualitative methods, or other techniques for studying rare events (King and Zeng, 2001). Trying to solve this problem by simply increasing the number of observations, warn Henry Brady and David Collier, 'may push scholars to compare cases that are not analytically equivalent' (2004: 11; also see Sartori, 1970).

As we have said repeatedly, testing hypotheses is a necessary part of social science. As a practical matter, however, the data limitations inherent in the IR field suggest that simplistic hypothesis testing will not yield as much progress as its practitioners believe. Instead, researchers must use theory to inform and guide the testing process.

Absence of explanation. As the well-known example of the democratic peace hypothesis illustrates, even well-confirmed empirical regularities do not provide an explanation for why they occur. A robust correlation still leaves us puzzled if we do not know why it happens and we tend to be skeptical of such findings until a convincing explanation — in other words, a theory — is given.[18]

Overemphasizing hypothesis testing also runs the risk of generating an ever-increasing body of empirical findings without identifying how they relate to each other. If one tests several hypotheses incorporating different independent variables and finds support for some but not others, the empirical results alone do not tell us why this is so. As David Dessler (1991: 340–341) notes, 'if theoretical integration implies a "tying together" of research findings, and not just a simple side-by-side listing of them … the heterogeneity of the independent variables is an obstacle to integration insofar as we lack a rationale for situating these quite different factors in relation to one another.'

For example, Reiter and Stam's *Democracies at War* tests a number of competing hypotheses about wartime performance, but, as Brooks (2003: 165) observes, it

> never offers a deductive argument for why some factors should be more powerful explanations than others.… Instead, Reiter and Stam test a diverse array of hypotheses … find empirical support for three, and then offer these findings as an explanation of democratic victory.

> Consequently, the argument about why democracy is such a sui generis phenomenon reads like a cumulation of disparate hypotheses. There is no true analytical engine driving the testing machine.

The recent literature on 'foreign imposed regime change' (FIRC) offers another example of this problem.[19] These works generally seek to determine whether FIRCs lead to positive outcomes (e.g. democracy, reduced danger of civil war, improved human rights performance, etc.). In some ways this literature is exemplary social science, especially given the difficulty of estimating the causal impact of a specific policy instrument such as military intervention on subsequent political and economic conditions.[20]

The best works in this genre have generated useful empirical generalizations, such as the finding that ousting a foreign government increases the risk of civil war, especially in poor or divided societies. But we still lack an overarching explanation of these findings. Thus, even in those fortunate circumstances where concepts are clear and the available data are good, a collection of confirmed hypotheses cannot by itself provide us with a coherent, integrated account of the phenomena in question. What is missing is both a compelling explanation for each individual hypothesis and a broader story about how they fit together.

Lack of cumulation. Advocates of hypothesis testing believe that this approach will produce a growing body of well-confirmed empirical findings and lead to a more rapid accumulation of knowledge about international affairs. The anticipated advance is not occurring, however, for several interrelated reasons.

For starters, the data on which many of these studies are based are imperfect, as previously discussed. Equally important, scholarship in the hypothesis-testing tradition often produces incompatible or non-comparable results because researchers examine the same issues using different data sets, focus on different time periods, define key terms in different ways, or employ different analytical techniques. As Beck et al. (2000: 21) note:

> statistical results appear to change from article to article and specification to specification. Any relationships usually are statistically weak, with wide confidence intervals, and they vary considerably with small changes in specification, index construction, and choice of data frame.

Unless a serious effort is made to reconcile these diverse studies and bring them into a common framework — which is the task of theory — there is little chance that knowledge will accumulate. If several published articles on a given topic all contain statistically significant but substantively different results and there is no theory to guide us, how do we decide which one to believe?

For example, in a generally positive review of the literature on interstate rivalries, John Vasquez and Christopher Leskiw (2001: 296–297) note that 'differences in operationalization led to different lists of [enduring] rivalries,' with different researchers being 'highly skeptical' of the definitions and lists used by others. As their essay makes clear, the definitional and methodological differences between competing studies led to an expanding set of empirical findings but did not produce a broader synthesis or a general explanation of the various positive and negative results. Instead, we get

generalizations of the following sort: 'Dyads that contend in territorial disputes have a greater probability of going to war than is expected by chance,' or '[Enduring] rivals have a greater probability of going to war than other dyads' (Vasquez and Leskiw, 2001: 308–309). But we still have little idea why.

The voluminous literature on ethnic and civil wars exhibits a similar lack of cumulation and for the same reasons. A recent survey of three decades of research found that prominent empirical studies often yield sharply different results, because they 'attach different interpretations to key variables,' 'differ in how they code civil wars,' rely on 'somewhat ad hoc empirical models,' and employ different explanatory variables, many of which are 'plausibly endogenous, biasing other estimates in unknown directions.' The authors conclude: 'ultimately, empirical work should aim to distinguish which of the competing theoretical mechanisms best explain the incidence, conduct, and nature of civil war, but this goal is still far from being realized' (Blattman and Miguel, 2010: 22–23).[21]

These examples suggest that simplistic hypothesis testing will not produce the cumulative progress its advocates expect. Indeed, these practices can even lead the same author to make contrasting claims in different articles, without providing an explanation for the different results.

For example, Jason Lyall (2009) finds that 'indiscriminate' violence by the Russian military reduced insurgent attacks in Chechnya. A second article found that counter-insurgent sweeps by local Chechen forces were more effective than sweeps conducted by Russian or mixed Russian–Chechen units, mainly because purely Chechen forces dealt with the local population in a more discriminating fashion (Lyall, 2010). Thus, in the first article, indiscriminate violence is the key to defeating Chechen insurgents, but in the second article, discriminate tactics are judged more effective.

Lyall and a co-author have published a third article arguing that reliance on more mechanized armies is 'associated with an increased probability of state defeat' in counter-insurgency campaigns (Lyall and Wilson, 2009: 67). This finding appears to be at odds with the claims in the first article, however, because the Russian army was highly mechanized and the indiscriminate tactics that supposedly worked in Chechnya consisted primarily of massive artillery bombardments. Each of these three studies may be defensible on its own and one can think of ways to reconcile the results, but together they create another puzzle to be explained rather than cumulative progress.

Last but not least, the belief that hypothesis testing alone will yield cumulative knowledge and useful predictions rests on the ancillary assumptions that the future will be more-or-less identical to the past and that results obtained in one context apply in other circumstances. In other words, we must assume that empirical generalizations uncovered by analyzing past data will be valid across space and time. This may be true in many instances, but we need theory to tell us when this is so. Because theories identify the causal connections between key variables as well as their boundary conditions, they explain when an observed relationship will persist, when a previously reliable generalization might weaken, and when a formerly weak association might become stronger.

To repeat: hypothesis testing is essential to social science and statistical analysis is a powerful tool when carried out properly. Furthermore, qualitative research can also suffer from poor data quality, selection bias, vague conceptualizations, lack of cumulation,

and other problems.[22] In short, our argument is not about privileging one set of methods over another. Rather, our argument is that the tendency for IR scholars to focus on methods and neglect theory is a step in the wrong direction. Thus far, this trend has not produced a large body of cumulative knowledge or a broad and enduring understanding of important international phenomena. Nor is it likely to in the future.

Why is IR headed in this direction?

Simplistic hypothesis testing may be more widespread today for reasons that are intellectually defensible, but its popularity has more to do with the professional incentives that academics now face.

To begin with, some might argue that there is not much new to say theoretically, especially at the level of grand theory. If theory development has reached the point of diminishing returns, then testing existing theories more carefully will yield greater insights. Until the next theoretical breakthrough, IR scholars should focus on exploring familiar puzzles with tried-and-true research approaches. In practice, this means testing hypotheses and devoting greater attention to middle-range theory.

This argument has some merit, as there is a substantial inventory of IR theories representing a wide range of perspectives. This fact does not justify the shift toward hypothesis testing, however, and especially the casual approach to theory that characterizes much of this work. As noted, simplistic hypothesis testing is not producing lots of cumulative knowledge. Furthermore, even if scholars are not trying to invent new theories or refine existing ones, their efforts to test hypotheses should be guided by a sophisticated understanding of theory, for reasons already discussed.

Moreover, one cannot be sure that a new grand theory or a powerful middle-range theory will not be created, especially given the emergence of new political conditions (e.g. unipolarity, globalization, etc.) that we want to understand. Nor should we forget that the existing body of grand theory still needs refinement, as the recurring debates among and within the isms illustrate. Many of the subjects covered by middle-range theory also remain poorly conceptualized, despite extensive efforts to test hypotheses relating to these topics.

Second, simplistic hypothesis testing may be more popular today because the availability of data and modern computer technology makes it easier to do. These developments may partly explain why the shift is occurring, but they do not justify it. We do have more software and more data at our fingertips, but much of the data we have are not very good despite impressive efforts to improve them.

We are agnostic about whether IR scholars will one day be able to use ‘big data’ sources and powerful data-mining techniques (such as those employed by companies like Google) to produce new and important insights. But even if these techniques eventually allow for more reliable predictions in some areas, they will do so by uncovering empirical patterns that need to be explained. Even when data are plentiful, theory is not easily dispensed with.

Third, the shift away from theory may reflect the impact of Gary King, Robert Keohane, and Sidney Verba’s (1994) book *Designing Social Inquiry*, which has been described as the ‘canonical text of the orthodox camp of political methodology’ (Schrodt,

2006: 335; see also Brady and Collier, 2004: 5; Yang, 2003). The book has been a staple of graduate-level methods courses because it offers an accessible template for doing social science. That template, notes Tim McKeown (1999: 162, 166), is based on 'the statistical worldview.' Moreover, *Designing Social Inquiry* fits squarely in the instrumentalist tradition: it 'privilege[s] observation and generalization at the expense of theory and explanation' (Johnson, 2006: 246). Insofar as this book became the ur text for how to do social science, it is not surprising that simplistic hypothesis testing also became more widespread.

Fourth, it is possible that this trend reflects the impact of the long debate on the democratic peace. It began with the empirical observation that 'democracies do not fight each other' (Doyle, 1983), and a cottage industry of subsequent large-*N* studies generally confirmed this claim. Yet there is still no convincing theory to account for this finding. Even without theory, it seemed, we could still learn new things about IR. Unfortunately, this literature may be a poor model for the field as a whole, because relationships as robust as the democratic peace are rare and searching for new ones at the expense of theory is likely to be counter productive (Reese, 2012).

Fifth, the expansion of PhD programs in IR encourages the shift toward hypothesis testing. It is hard for any graduate program to produce top-notch theorists because theoretical fertility depends primarily on individual creativity and imagination. No one knows how to teach people to be creative, however, and no one has yet identified a program of study that would enable a department to crank out brilliant theorists en masse.[23] By contrast, almost anyone with modest mathematical abilities can be taught the basic techniques of hypothesis testing and produce competent research. Similarly, teaching students about research design, process-tracing, and historical interpretation can help them do better qualitative research, but it will not turn someone lacking imagination into an accomplished theorist.

Moreover, because graduate programs are reducing the time students take to complete their degrees, teaching a set of tools that enable them to produce a competent thesis quickly has become the norm. Developing or refining theory is more time-consuming and riskier as it requires deeper immersion in the subject matter and the necessary flash of inspiration may never occur. Once a graduate program is committed to getting lots of PhD students out the door on schedule, it has a powerful incentive to emphasize simplistic hypothesis testing. In addition, piling on more and more methods courses (whether quantitative or qualitative) while compressing the time to degree inevitably crowds out courses on theory and on the substance of IR, leaving students ill-equipped to think in creative and fruitful ways about the field's core issues.

Sixth, privileging hypothesis testing creates more demand for empirical work and thus for additional researchers. As hypothesis testing becomes ascendant, the field will generate more and more studies without resolving much. Confirming the work of other researchers garners little attention or prestige, so scholars naturally focus their efforts on producing novel findings and challenging prior work. Generating novel results is easy, however, when the relevant variables are defined in different ways, data quality is poor, and the hypotheses being tested are loosely tied to theory. As discussed above, these problems typify much of the hypothesis testing that takes place in IR. Under these conditions, regression coefficients 'can bounce around like a box of gerbils on

methamphetamines. This is great for generating large bodies of statistical literature … but not so great at ever coming to a conclusion' (Schrodt, 2006: 337). Because research rarely cumulates, there will always be new studies to perform, thereby generating a self-perpetuating demand for scholars to perform them. The more hypothesis testers we produce, it seems, the more hypothesis testers we need.

Lastly, the appeal of simplistic hypothesis testing reflects the professionalization of academia. Like other professions, academic disciplines strive to safeguard their autonomy and maximize the prestige and material benefits accruing to their members. One way to do this is to convince outsiders that the profession has specialized expertise. Thus, professions have powerful incentives to employ esoteric terminology and arcane techniques that make it difficult to evaluate what its members are saying. This tendency is apparent in the hypothesis-testing literature, as even a cursory reading of IR journals reveals.

Over time, professions also tend to adopt simple and impersonal ways to evaluate members. In the academy, this tendency leads to heavy reliance on 'objective' criteria — such as citation counts — in hiring and promotion decisions. In some cases, department members and university administrators might think that they do not have to read a scholar's work and form an independent opinion of its quality. Instead, they can simply calculate the individual's 'h-index' (Hirsch, 2005) and make personnel decisions on that basis.[24]

These tendencies encourage scholars to move away from theory and toward hypothesis testing. Such works often employ statistical techniques that require a significant investment of time to master. Those who lack such training cannot easily criticize these works, and some members of a department may not be able to tell if a colleague's research is truly significant. They will have to rely on appraisals from scholars who do the same kind of work or on some other measure of merit. When you do not understand someone's work but you still have to judge it, you will be tempted to ask 'How many articles has she published?' or 'How many other people cite his work?' In this way, reliance on esoteric terminology and arcane techniques inhibits others from directly evaluating scholarly merit.

Obviously, the more universities rely on 'objective' measures to evaluate scholars, the greater the incentive to adopt a research strategy that maximizes the number of publications one can produce quickly. These incentives are apparent to today's hyper-professionalized graduate students, who worry that getting a job requires them to publish as soon and often as possible. They are understandably drawn to simplistic hypothesis testing, which allows them to take a data set and start cranking out articles, either by varying the research questions slightly, employing a series of different models, or using new estimation techniques.[25]

These same incentives encourage scholars to tread well-worn research paths, making it more likely that lots of other people will read and cite their work. Unfortunately, such herd-like behavior reinforces scholarly fads and discourages bolder and more original work (Jervis, 1976). As Vinod Aggarwal (2010: 895) notes:

> Simply put, quantitative research using data sets that address narrow issues provides a risk-averse … path to tenure. MPUs (minimum publishable units) rule the day. Why risk conceptual

> or ontological innovation that might not be well received, when plodding along with marginal contributions will raise one's point count? The result is worship at the Social Science Citation Index altar … that does little to foster innovation and creativity.

The rise of simplistic hypothesis testing and the declining interest in theory has also increased the gulf between academia and the policy world. As discussed above, theory is essential for understanding a complex reality, for formulating policy responses, and for policy evaluation. For example, how one thinks about dealing with a rising China depends first and foremost on one's broad perspective on world politics. Realist theories suggest one set of responses; liberal or constructivist theories offer quite different policy recommendations (Fravel, 2010; Liu, 2010). Creating and refining theories is an activity that academics are uniquely well positioned to do. When academics lose interest in theory, therefore, they relinquish one of their most potent weapons for influencing critical policy debates.

This situation may not trouble those hypothesis testers who are primarily concerned with professional advancement. What matters to them is one's citation count, not helping outsiders understand important policy issues. As we have seen, the hypothesis-testing culture has produced little reliable or useful knowledge, and its esoteric jargon and arcane methods have made IR scholarship less accessible to policymakers, informed elites, and the public at large. Moreover, the emergence of an extensive think-tank community in Washington, London, and other world capitals has made policymakers less dependent on IR scholars at precisely the moment when these same scholars have less to contribute. Taken together, these trends run the risk of making IR largely irrelevant to understanding and solving important real-world problems.

Can anything be done?

IR is a conceptually complex and diverse field where reliable data are hard to come by. These features require scholars to rely more heavily on theory than their counterparts in other areas of social science. It follows that the field should privilege theory, as it once did. Instead, IR is headed in the opposite direction.

IR scholars should test hypotheses, of course, but in ways that are guided by a well-specified theory. They should also focus considerable attention on refining existing theories and developing new ones. In particular, greater effort should be devoted to investigating the causal mechanisms implied by different theories. A single article that advances a new theory or makes sense of a body of disparate findings will be more valuable than dozens of empirical studies with short shelf-lives.

Some may argue we have overstated the problem and that the field is addressing the shortcomings we identify. Some scholars now focus on micro-level questions for which more reliable data are available (Kalyvas, 2008), while others seek to minimize the need for theory by using natural, field, or laboratory experiments to provide exogenous variation (Tomz and Weeks, 2013; Yanigazawa-Drott, 2010). Yet in the absence of well-developed theory, we have no way of knowing whether the results from individual experiments are generalizable.[26] Moreover, focusing on issues where experiments are feasible is likely to direct the IR field toward questions of lesser substantive importance. A few

scholars are exploring new methods for studying causal mechanisms (Imai et al., 2011) or developing other statistical techniques to deal with missing data or other problems of inference (Beck et al., 2000; King and Zeng, 2001). What remains to be seen is whether these efforts can generate new and important insights about the substance of IR. To date, the results have been meager.[27]

What might restore theory to its proper place? Academic disciplines are socially constructed and self-policing; if enough IR scholars thought that the present approach was not working, they could reverse the present trajectory. But such an epiphany is unlikely. Powerful professional incentives encourage an emphasis on simplistic hypothesis testing, and the rise of think tanks and consulting firms has reduced demand for academic scholarship on policy issues. IR scholars are less inclined to develop and refine theories or perform theory-guided empirical tests, therefore, and we are not optimistic that this situation will change.

To be sure, a few university administrators may not like the direction in which IR is moving and may try to encourage departments to move away from the 'dreary hypothesis-testing framework.' Foundations that fund research might recognize the problems we identify and offer to support more theoretically or policy-oriented work. But academic disciplines usually resist outside interference and change would have to occur in many departments, not just one or two.

Finally, external events might encourage theoretical innovation and policy engagement, especially if citizens and policymakers face unexpected challenges and need new theories to grasp them. Unfortunately, there is little evidence that any of these potential catalysts for change will push IR back toward theory.

What might be done to encourage the shift we advocate? Emphasizing quality over quantity in a scholar's portfolio might help. If faculty understood that hiring and promotion depended on evaluating only three or four publications, they might focus on producing scholarship of greater significance instead of maximizing the total number of peer-reviewed articles. This would be a partial remedy at best, however, because those involved in personnel decisions would still be aware of a candidate's full inventory of publications and unlikely to ignore it completely. Even if this norm were adopted, its impact would be modest.

In our view, therefore, the present emphasis on hypothesis testing is unlikely to change. Nevertheless, scholars in the field are free agents, and perhaps a critical mass of them will see the light and restore theory to its proper place in the study of international politics.

Conclusion

The study of IR should be approached with humility. There is no single theory that makes understanding world politics easy, no magic methodological bullet that yields robust results without effort, and no search engine that provides mountains of useful and reliable data on every question that interests us. We therefore favor a diverse intellectual community where different theories and research traditions coexist. Given how little we know, and how little we know about how to learn more, overinvesting in any particular approach seems unwise. As Schrodt (2006: 336) wisely observes, 'we need all the help we can get to figure out this whacko world.'

What matters most, however, is whether we create more powerful theories to explain key features of IR. Without good theories, we cannot trust our empirical findings, whether quantitative or qualitative in nature. Unless we have theories to make sense of them, we cannot even keep track of all the hypotheses that scholars keep piling up. There are many roads to better theory, but that should be our ultimate destination.

Acknowledgements

We are deeply indebted to the following individuals for comments, suggestions, or helpful discussions on this article: Andrew Abbott, Andrew Bennett, Bear Braumoeller, Thomas Christensen, Dara Kay Cohen, Alexandre Debs, Michael Desch, John Duffield, Jeffrey Friedman, Charles Glaser, Hein Goemans, James Johnson, Burak Kadercan, Austin Knuppe, Paul MacDonald, Nuno Monteiro, Michael Reese, Dan Reiter, Marie-Eve Reny, Michael Rowley, Allan Stam, Paul Staniland, Michael Weintraub, David Yanigazawa-Drott, Richard Zeckhauser, and Yuri Zhukov. We are also grateful for comments received at seminars at Harvard's Belfer Center for Science and International Affairs, the Georgetown University International Theory and Research Seminar, and the Notre Dame International Security Program.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Notes

1. Four different TRIP surveys have asked IR scholars to identify the 'best,' 'most interesting,' or 'most influential' work in the field. There is considerable overlap in the responses and well-known theorists dominate the lists (see Maliniak et al., 2007: 17–19; 2012: 48–50; Peterson et al., 2005: 19–21).
2. The authors of the TRIP surveys note that there has been a 'dramatic decline of atheoretic work from 47% in 1980 to 7% in 2006' (Maliniak et al., 2011: 445). This finding reflects the fact that almost all contemporary IR scholars pay some homage to theory in their work. Our point, however, is that theory usually plays a minor role.
3. This is sometimes true for scholars who favor qualitative methods as well (see Bennett and Elman, 2007; Moravcsik, 2010).
4. Despite similar names, scientific realism and the realist approach to international relations are wholly distinct. The former is a school of thought in epistemology; the latter is an approach to international politics. Thus, one could be a 'scientific realist' and reject realism in IR, or vice versa. On the differences between scientific realism and instrumentalism, see MacDonald (2003: 551–565; also see Chakravartty, 2011; Clarke and Primo, 2007: 741–753; George and Bennett, 2004: ch. 7; Johnson, 2010).
5. Achen and Snidal (1989: 164) illustrate instrumentalism in their characterization of deterrence theory: 'Rational deterrence theory is agnostic about the actual calculations decision makers undertake. It holds that they will act as if they solve certain mathematical problems whether or not they actually solve them. Just as Steffi Graf plays tennis as if she did rapid computations in Newtonian physics … so rational deterrence theory predicts that decision makers will decide whether to go to war as if they did expected utility calculations. But they need not actually perform them.'
6. Some scholars maintain that formal theory is especially well suited for producing logically consistent arguments (see Bueno de Mesquita and Morrow, 1999: 56–60). Yet they admit

that non-formal theories can also be logically consistent and the use of mathematics does not prevent logical mistakes. Indeed, complicated mathematical proofs can be less accessible and more difficult to verify. As Melvyn Nathanson (2009: 9) observes: 'the more elementary the proof, the easier it is to check and the more reliable is its verification.' And we would argue that creativity and originality are more important than mere logical consistency (see Walt, 1999: 116–118).

7. Although measuring covariation is usually identified with large-*N* research, it is also possible with qualitative research or case studies (see King et al., 1994).
8. On causal mechanisms, see George and Bennett (2004: ch. 10), Hedstrom and Ylikoski (2010), Johnson (2010), Mahoney (2001), Waldner (2007), and Van Evera (1997: 64–67.)
9. For example, Thomas Schelling's influential ideas about compellence do not fare well when tested empirically. Nonetheless, scholars such as Wallace J. Thies and Robert A. Pape began with Schelling's ideas when fashioning their own theories of military coercion (see Pape, 1996; Schelling, 1966; Thies, 1980)
10. Quoted in Dessler (1991: 349).
11. Theory is not necessary for identifying puzzles that can lead scholars to invent new hypotheses. Sometimes, researchers observe something in the data that no theory can explain, so they try to come up with a story to account for it. Existing theories help scholars identify these anomalies, however, whenever what they are observing runs counter to their beliefs about how the world works. Scholars can also use hypothesis tests to determine which of two competing theories is most promising, even if the theories themselves are not well developed. A good example of this sort of work is Shapiro and Weidmann (2012).
12. This approach entails a shift away from constructing multivariate models that include all the relevant variables needed to account for a particular phenomenon (but no more), and toward models intended to assess the relative impact of different explanatory variables. As James L. Ray (2003: 3) notes:

    > general models aimed at the best fit for the model as a whole seem to have given way almost entirely to models whose basic purpose is to evaluate the impact of one key factor. Variables beyond that one key factor are added almost entirely for the purpose of providing a more sophisticated, thorough, and rigorous evaluation of a key hypothesis in question …. Most specifically … control variables are added to multivariate models in order to see whether the relationship of special interest persists.

13. This problem is compounded if researchers discard models that do not 'work' and report only those results that reach some canonical level of statistical significance. As Philip Schrodt (2006: 337) warns:

    > the ubiquity of exploratory statistical research has rendered the traditional frequentist significance test all but meaningless. Alternative models can now be tested with a few clicks of a mouse and a few seconds of computation. … Virtually all published research now reports only the final tip of an iceberg of dozens of unpublished alternative formulations. In principle, significance tests could be adjusted to account for this, in practice they are not.

14. See Fearon (1994), Huth and Russett (1984), and Huth (1988, 1990).
15. Quoting Hubert Blalock, James L. Ray (2003) points out that:

    > if one adds an intervening variable to a multivariate model, and this modification eliminates the statistical association between the original key explanatory factor being evaluated and the outcome variable, then one has engaged in 'interpretation' of that relationship. Such

'interpretation' does not make the original relationship in question less interesting. On the contrary, 'through interpretation one … is merely making it more plausible by finding the intermediate links.' This is a fundamentally different situation than that resulting from the addition of a potential confounding variable to a model that eliminates the correlation between the original independent and dependent variables. In that case, one is discovering that there is something radically wrong with the notion that X causes Y.

Also see Seawright (2010: 250–251).

16. For example, Alastair Iain Johnston (2012: 57) examined the coding of China cases from 1992 to 2001 in the MID data set and found errors in 12 out of 28.
17. As Van Evera (2009: 7) notes, 'We know a great deal about the twenty most data-rich instances of the outbreak of war. … But the data thin out fast as we move down the list from data-rich to data-poor wars.' Focusing on well-documented cases can be problematic, however, if they are not a random sample of the larger universe.
18. Some scholars argue that the absence of war between democracies is a statistical artifact or due to great-power politics or some other factor (Farber and Gowa, 1995; Gibler, 2007; Gowa, 1999). If true, then there is no such thing as a *democratic* peace. As always, the meaning of any empirical finding depends on theoretical interpretation.
19. Representative works include Pickering and Peceny (2006), Peic and Reiter (2011), Downes and Monten (2013), and Downes (2010).
20. Such studies have to contend with powerful selection effects and potential omitted variable bias, which is why some scholars working in this area have relied on matching techniques to strengthen the validity of their results.
21. Elisabeth Jean Wood (2003: 251) agrees: 'the emergence and course of identity conflicts is extremely difficult to trace statistically for a variety of reasons. As a result, the relevant findings are often contradictory.' Hegre and Sambanis (2006: 532) conducted a global sensitivity analysis of the civil war literature and conclude, 'no study to date has produced a clear theoretical justification for the model used in econometric tests. We do not know *the* model of civil war.' Also see Cederman et al. (2010: 90–91).
22. For example, Alexander George and Richard Smoke's prize-winning *Deterrence in American Foreign Policy: Theory and Practice* (1974) addresses only cases of deterrence failure (selection bias), offers 'contingent empirical generalizations' rather than genuine theory, and provides little cumulative knowledge about when deterrence will succeed or fail.
23. This point applies to both formal and non-formal theory. One can teach students the basic techniques of formal modeling, but not all will become creative formal theorists.
24. Scott and Light (2004) provide a provocative critique of this general approach.
25. As Achen (2002: 442) notes:

> Empirical work, the way too many political scientists do it, is indeed relatively easy. Gather the data, run the regression/MLE with the usual linear list of control variables, report the significance tests, and announce that one's pet variable 'passed.' This dreary hypothesis-testing framework is sometimes seized upon by beginners. Being purely mechanical, it saves a great deal of thinking and anxiety, and cannot help being popular. But obviously, it has to go. Our best empirical generalizations do not derive from that kind of work.

Nor, we might add, do our best theories.

26. Experimental and quasi-experimental work in development economics suffers from a similar deficiency (see Deaton, 2010).
27. This problem appears to be prevalent in both sociology and economics. As sociologist Aage Sorensen notes, "quantitative sociology remains very theory-poor. In fact, the mainstream

has regressed rather than progressed" (quoted in Mahoney 2001: 582). Inattention to theory also leads to questionable inferences in empirical economic research (Wolpin 2013; also see Hamermesh 2013).

References

Achen CH (2002) Toward a new political methodology: Microfoundations and ART. *Annual Review of Political Science* 5: 423–450.

Achen CH (2005) Let's put garbage-can regressions and garbage-can probits where they belong. *Conflict Management and Peace Science* 22(3): 327–339.

Achen CH and Snidal D (1989) Rational deterrence theory and comparative case studies. *World Politics* 41(2): 143–169.

Aggarwal VK (2010) I don't get no respect: The travails of IPE. *International Studies Quarterly* 54(3): 893–895.

Beck N, King G and Zeng L (2000) Improving quantitative studies of international conflict: A conjecture. *American Political Science Review* 94(1): 21–35.

Bennett A and Elman C (2007) Case study methods in the International Relations subfield. *Comparative Political Studies* 40(2): 170–195.

Bennett A, Barth A and Rutherford KR (2003) Do we preach what we practice? A survey of methods in political science journals and curricula. *PS: Political Science & Politics* 36(3): 373–378.

Blattman C and Miguel E (2010) Civil war. *Journal of Economic Literature* 48(1): 3–57.

Brady HE and Collier D (eds) (2004) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. New York: Rowman & Littlefield.

Braumoeller BF and Sartori AE (2004) The promise and perils of statistics in International Relations. In: Sprinz D and Wolinsky Y (eds) *Models, Numbers, and Cases: Methods for Studying International Relations*. Ann Arbor, MI: University of Michigan Press, 129–151.

Brooks RA (2003) Making military might: Why do states fail and succeed? A review essay. *International Security* 28(2): 149–191.

Brooks S (2007) *Producing Security: Multinational Corporations, Globalization, and the Changing Calculus of Conflict*. Princeton, NJ: Princeton University Press.

Bueno de Mesquita B and Morrow J (1999) Sorting through the wealth of notions. *International Security* 24(2): 56–73.

Cederman LE, Wimmer A and Min B (2010) Why do ethnic groups rebel? New data and analysis. *World Politics* 62(1): 87–119.

Chakravartty A (2011) Scientific realism. In: Zalta E (ed.) *The Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/archives/sum2011/entries/scientific-realism/ (accessed 24 October 2012).

Chen HT (1990) *Theory Driven Evaluations*. Thousand Oaks, CA: Sage Publications.

Clarke KA and Primo DM (2007) Modernizing political science: A model-based approach. *Perspectives on Politics* 5(4): 741–753.

Cohen BJ (2010) Are IPE journals becoming boring? *International Studies Quarterly* 54(3): 887–891.

Deaton A (2010) Instruments, randomization, and learning about development. *Journal of Economic Literature* 48(2): 424–455.

Dessler D (1991) Beyond correlations: Toward a causal theory of war. *International Studies Quarterly* 35(3): 337–355.

Downes AB (2010) Catastrophic success: Foreign-imposed regime change and civil war. In: *American Political Science Association Conference*, Washington, DC, 2–5 September 2010.

Downes AB and Monten J (2013) Forced to be free? Why foreign-imposed regime change rarely leads to democratization. *International Security* 37(4): 90–131.

Downes AB and Sechser T (2012) The illusion of democratic credibility. *International Organization* 66(3): 457–489.
Doyle MW (1983) Kant, liberal legacies and foreign affairs. *Philosophy and Public Affairs* 12(3): 205–235.
Farber HS and Gowa J (1995) Polities and peace. *International Security* 20(2): 123–146.
Fearon JD (1994) Signaling versus the balance of power and interests. *Journal of Conflict Resolution* 38(2): 236–269.
Fravel MT (2010) International relations theory and China's rise: Assessing China's potential for territorial expansion. *International Studies Review* 12(4): 505–532.
Friedman M (1953) *Essays in Positive Economics*. Chicago, IL: University of Chicago Press.
George AL and Bennett A (2004) *Case Studies and Theory Development in the Social Sciences*. Cambridge, MA: Belfer Center for International Affairs, Harvard University.
George AL and Smoke R (1974) *Deterrence in American Foreign Policy: Theory and Practice*. New York: Columbia University Press.
Gibler D (2007) Bordering on peace: Democracy, territorial issues, and conflict. *International Studies Quarterly* 51(3): 509–532.
Gowa J (1999) *Ballots and Bullets: The Elusive Democratic Peace*. Princeton, NJ: Princeton University Press.
Hamermesh D (2013) Six decades of top economics publishing: Who and how? *Journal of Economic Literature* 51(1): 1–11.
Hedstrom P and Ylikoski P (2010) Causal mechanisms in the social sciences. *Annual Review of Sociology* 36: 49–67.
Hegre H and Sambanis N (2006) Sensitivity analysis of empirical results on civil war onset. *Journal of Conflict Resolution* 50(4): 508–535.
Hirsch JE (2005) An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*. Available at: http://arxiv.org/PS_cache/physics/pdf/0508/0508025v5.pdf (accessed 24 October 2012).
Huth P (1988) *Extended Deterrence and the Prevention of War*. New Haven, CT: Yale University Press.
Huth P (1990) The extended deterrent value of nuclear weapons. *Journal of Conflict Resolution* 34(2): 270–290.
Huth P and Russett B (1984) What makes deterrence work? Cases from 1900 to 1980. *World Politics* 36(4): 496–526.
Ikenberry J, Mastanduno M and Wohlforth W (eds) (2011) *International Relations Theory and the Consequnces of Unipolarity*. Cambridge: Cambridge University Press.
Imai K, Keele L, Tingley D et al. (2011) Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* 105(4): 765–789.
Jervis R (1976) Cumulation, correlations and woozles. In: Rosenau JN (ed.) *In Search of Global Patterns*. New York: Free Press, 181–185.
Johnson J (2006) Consequences of positivism: A pragmatist assessment. *Comparative Political Studies* 39(2): 224–252.
Johnson J (2010) What rationality assumption? Or, how 'positive political theory' rests on a mistake. *Political Studies* 58(2): 282–299.
Johnston AI (2012) What (if anything) does East Asia tell us about International Relations theory? *Annual Review of Political Science* 15: 53–78.
Jordan R, Maliniak D, Oakes A et al. (2009) One discipline or many? TRIP survey of international relations faculty in ten countries. *Report, The Institute for the Theory and Practice of International Relations, The College of William and Mary, VA*, February.

Kalyvas SN (2008) Promises and pitfalls of an emerging research program: The microdynamics of civil war. In: Kalyvas SN, Shapiro I and Masoud T (eds) *Order, Conflict, Violence*. Cambridge: Cambridge University Press, 397–421.

Kaplan F (1983) *The Wizards of Armageddon*. New York: Simon and Schuster.

King G and Zeng L (2001) Explaining rare events in international relations. *International Organization* 55(3): 693–715.

King G, Keohane RO and Verba S (1994) *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.

Lake DA (2011) Why 'isms' are evil: Theory, epistemology, and academic sects as impediments to understanding and progress. *International Studies Quarterly* 55(2): 465–480.

Layne C (1994) Kant or cant: The myth of the democratic peace. *International Security* 19(2): 5–49.

Liu Q (2010) China's rise and regional strategy: Power, interdependence and identity. *Journal of Cambridge Studies* 5(4): 76–92.

Lyall J (2009) Does indiscriminate violence incite insurgent attacks? Evidence from Chechnya. *Journal of Conflict Resolution* 53(3): 331–362.

Lyall J (2010) Are coethnics more effective counterinsurgents? Evidence from the second Chechen war. *American Political Science Review* 104(1): 1–20.

Lyall J and Wilson I III (2009) Rage against the machines: Explaining outcomes in counterinsurgency wars. *International Organization* 63(1): 67–106.

Lynn-Jones SM (1995) Offense-Defense theory and its critics. *Security Studies* 4(4): 660–691.

MacDonald PK (2003) Useful fiction or miracle maker: The competing epistemological foundations of rational choice theory. *American Political Science Review* 97(4): 551–565.

McKeown TJ (1999) Case studies and the statistical worldview: Review of King, Keohane, and Verba's *Designing Social Inquiry: Scientific Inference in Qualitative Research*. *International Organization* 53(1): 161–190.

Mahoney J (2001) Beyond correlational analysis: Recent innovations in theory and method. *Sociological Forum* 16(3): 575–593.

Maliniak D, Oakes A, Peterson S et al. (2007) The view from the ivory tower: TRIP survey of international relations faculty in the United States and Canada. *Report, The Institute for the Theory and Practice of International Relations, The College of William and Mary, VA*, February.

Maliniak D, Oakes A, Peterson S et al. (2011) International relations in the US academy. *International Studies Quarterly* 55(2): 437–464.

Maliniak D, Peterson S and Tierney MJ (2012) TRIP around the world: Teaching, research, and policy views of international relations faculty in 20 countries. *Report, The Institute for the Theory and Practice of International Relations, The College of William and Mary, VA*, May.

Mansfield ED and Snyder J (2007) *Electing to Fight: Why Emerging Democracies Go to War*. Cambridge, MA: MIT Press.

Monteiro NP (2011/ 2012) Unrest assured: Why unipolarity is not peaceful. *International Security* 36(3): 9–40.

Moravcsik A (2010) Active citation: A precondition for replicable qualitative research. *PS: Political Science and Politics* 43(1): 29–35.

Nathanson M (2009) Desperately seeking mathematical proof. *The Mathematical Intelligencer* 31(2): 8–10.

Oatley T (2011) The reductionist gamble: Open economy politics in the global economy. *International Organization* 65(2): 311–341.

Olson M (1965) *The Logic of Collective Action: Public Goods and the Theory of Groups*. Cambridge, MA: Harvard University Press.

Ostrom E (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Pape RA (1996) *Bombing to Win: Airpower and Coercion in War*. Ithaca, NY: Cornell University Press.

Peic G and Reiter D (2011) Foreign imposed regime change, state power, and civil war onset, 1920–2004. *British Journal of Political Science* 41(3): 453–475.

Peterson S, Tierney MJ and Maliniak D (2005) Teaching and research practices, views on the discipline, and policy attitudes of international relations faculty at US colleges and universities. *Report, Program on the Theory and Practice of International Relations, The College of William and Mary, VA*, August.

Pickering J and Peceny M (2006) Forging democracy at gunpoint. *International Studies Quarterly* 50(3): 539–560.

Putnam H (1975) *Mathematics, Matter and Method: Philosophical Papers, Volume 1*. London: Cambridge University Press.

Ray JL (2003) Explaining interstate conflict and war: What should be controlled for? *Conflict Management and Peace Science* 20(1): 1–31.

Reese MJ (2012) Personal communication.

Reiter D and Stam AC (2002) *Democracies at War*. Princeton, NJ: Princeton University Press.

Sartori G (1970) Concept misformation in comparative politics. *American Political Science Review* 64(4): 1033–1053.

Schelling TC (1966) *Arms and Influence*. New Haven, CT: Yale University Press.

Schrodt P (2006) Beyond the linear frequentist orthodoxy. *Political Analysis* 14(3): 335–339.

Schrodt P (2010) Seven deadly sins of contemporary quantitative political analysis. Paper prepared for the annual meeting of the American Political Science Association, Washington, DC.

Scott J and Light M (2004) The misuse of numbers: Audits, quantification, and the obfuscation of politics. In: Purdy J, Kronman AT and Farrar C (eds) *Democratic Vistas: Reflections on the Life of American Democracy*. New Haven, CT: Yale University Press.

Seawright J (2010) Regression-based inference: A case study in failed causal assessment. In: Brady HE and Collier D (eds) *Rethinking Social Inquiry* (2nd edn). New York: Rowman and Littlefield.

Shapiro JN and Weidmann NB (2012) Is the phone mightier than the sword? Cell phones and insurgent violence in Iraq. *Princeton University* working paper.

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

Snyder JL (1984/ 1985) Richness, rigor, and relevance in the study of Soviet foreign policy. *International Security* 9(3): 89–108.

Tapp C, Burkle FM, Wilson W, et al. (2008) Iraq war mortality estimates: A systematic review. *Conflict and Health* 2:1.

Thies WJ (1980) *When Governments Collide: Coercion and Diplomacy in the Vietnam Conflict, 1964–1968*. Berkeley, CA: University of California Press.

Tomz M and Weeks JL (2013) Public opinion and the democratic peace. *American Political Science Review* 107(3): forthcoming.

Trachtenberg M (2006) *The Craft of International History: A Guide to Method*. Princeton, NJ: Princeton University Press.

Van Evera S (1984) The cult of the offensive and the origins of the First World War. *International Security* 9(1): 58–107.

Van Evera S (1997) *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Van Evera S (2009) Trends in political science and the future of security studies. 2009–10 Annual Report, Security Studies Program, Massachusetts Institute of Technology, MA.
Vasquez J and Leskiw CS (2001) The origins and war proneness of interstate rivalries. *Annual Review of Political Science* 4: 295–316.
Von Clausewitz C (1976) *On War*. Translated and edited by Howard M and Paret P. Princeton, NJ: Princeton University Press.
Waldner D (2007) Transforming inferences into explanations: Lessons from the study of mass extinctions. In: Lebow RN and Lichbach MI (eds) *Theory and Evidence in Comparative Politics and International Relations*. New York: Palgrave Macmillan.
Walt SM (1999) A model disagreement. *International Security* 24(2): 115–130.
Weart S (1998) *Never at War: Why Democracies Will Not Fight One Another*. New Haven, CT: Yale University Press.
Weaver C (ed) (2009) Special issue: Not so quiet on the western front: The American school of IPE. *Review of International Political Economy* 16(1).
Wohlforth WC (1999) The stability of a unipolar world. *International Security* 24(2): 5–41.
Wohlstetter A (1959) The delicate balance of terror. *Foreign Affairs* 37(2): 211–234.
Wolpin K (2013) *The limits of inference without theory*. Cambridge, MA: MIT Press.
Wood EJ (2003) Civil wars: What we don't know. *Global Governance* 9(2): 247–260.
Yang DD (2003) Qualitative methods syllabi. *Qualitative Methods Newsletter* 1(1): 28–30.
Yanigazawa-Drott D (2010) Propaganda and conflict: Theory and evidence from the Rwanda genocide. Working Paper, Harvard University.

Author biographies

John J. Mearsheimer is R. Wendell Harrison Distinguished Service Professor of Political Science at the University of Chicago, USA.

Stephen M. Walt is Robert and Renee Belfer Professor of International Affairs at the John F. Kennedy School of Government, Harvard University, USA.