The European Journal of International Law Vol. 28 no. 4

Experimenting with International Law

Jeffrey L. Dunoff[*] and Mark A. Pollack[**]

Abstract

A growing body of experimental research has begun to explore the causal mechanisms through which international law impacts behaviour. International legal scholars, however, are still in the early stages of adopting experimental methods. Indeed, Yahli Shereshevsky and Tom Noah's article is one of the first experimental studies to appear in the European Journal of International Law. *Its publication thus provides an opportunity to reflect not only on this pioneering work but also on the broader 'experimental turn' in the study of international law. To do so, we begin by motivating the experimental turn, which we argue reflects both a methodological shift from observational studies towards the increasing use of experiments and a theoretical shift from rational choice towards cognitive psychology and behavioural economics. Second, we engage in a critical reading of Shereshevsky and Noah's study of the impact of preparatory materials on treaty interpretation. Applying the dual criteria of internal and external validity, we assess the strengths and weaknesses of Shereshevsky and Noah's study. We conclude that experiments promise to extend our knowledge of international law and are likely to become increasingly influential in scholarly and policy debates. Hence, all international lawyers have an urgent interest in becoming knowledgeable and critical consumers of experimental research.*

A rapidly growing body of experimental research, conducted primarily by political scientists, has begun to identify the varied causal mechanisms through which international law impacts behaviour. International legal scholars, however, are still in the early stages of adopting (and adapting) experimental methods. Indeed, Yahli Shereshevsky and Tom Noah's article is one of the first experimental studies to appear in the pages of the *European Journal of International Law* (*EJIL*).[1] Its publication thus

[*] Laura H. Carnell Professor of Law, Beasley School of Law, Temple University, Philadelphia, PA, USA. Email: dunoffj@temple.edu.

[**] Professor of Political Science and Law, Jean Monnet Chair, Temple University, Philadelphia, PA, USA. Email: mark.pollack@temple.edu.

We are grateful to Sarah Bush, Adam Chilton, David Hoffman and Katerina Linos for thoughtful comments on earlier drafts.

[1] Shereshevsky and Noah, 'Does Exposure to Preparatory Work Affect Treaty Interpretation? An Experimental Study on International Law Students and Experts', 28 *European Journal of International Law* (*EJIL*)(2017) 1287. The first experimental study in the journal, to our knowledge, was published earlier this year: Puig and Strezhnev, 'The David Effect and ISDS', 28 *EJIL* (2017), 731.

provides a timely opportunity to reflect not only on this pioneering work but also on the broader 'experimental turn' in the study of international law.

Our aims in this brief article are three-fold. First, as a prelude to assessing Shereshevsky and Noah's study, we step back and situate their research in the context of the larger turn to experiments. The rise of experimental methods in international law, we argue in Part 1, reflects both a methodological shift from observational studies towards the increasing use of experiments and a theoretical shift from neo-classical rational choice assumptions towards a behavioural analysis based on empirical findings from cognitive psychology and behavioural economics. Because both of these shifts are quite recent in the study of international law, we explore each in turn, and we provide a very brief guide to the assessment of experimental studies, emphasizing the criteria of internal and external validity. As explained more fully below, a typical internal validity question is whether variations in a dependent variable can be confidently attributed to a hypothesized independent variable within an experimental setting, whereas a typical external validity question asks whether behaviours observed in experimental settings can be generalized to other real-life settings.

Second, having laid this groundwork, we turn in Part 2 to Shereshevsky and Noah's study of the impact of preparatory materials on treaty interpretation. Applying the dual criteria of internal and external validity, we assess the many strengths of Shereshevsky and Noah's study and note the ways in which future studies can build on it to increase our confidence in both the internal and external validity of its findings.

Third and finally, our overarching goal is to spark a larger dialogue over the promise – and the perils – of 'experimenting with international law'. Experiments, we argue, can deepen and extend our knowledge of international law's workings and impacts and, for this reason, are likely to become increasingly influential in scholarly and policy debates. Hence, all international lawyers – even those with little interest in conducting experimental research themselves – have an urgent interest in becoming knowledgeable and critical consumers of experimental research.

1 Why Experiments? Methodological and Theoretical Shifts

EJIL readers will be familiar with the decades-long rise of empirical legal studies, which involves the systematic analysis of qualitative or quantitative data, often using statistical techniques on large data sets.[2] The use of experiments, a particular form of empirical investigation, in the study of international law is of a much more recent vintage, however, with the first explicitly experimental studies of international law appearing only in the past half-decade. This development invites an examination of the reasons for the adoption of experimental methodologies in both international relations and international law, the criteria used to judge the validity of experimental studies and

[2] Shaffer and Ginsburg, 'The Empirical Turn in International Legal Scholarship', 106 *American Journal of International Law* (*AJIL*) (2012) 1. See also Hafner-Burton, Victor and Lupu, 'Political Science Research on International Law: The State of the Field', 106 *AJIL* (2012) 47.

an overall assessment of the advantages and disadvantages of experimental methods for students of international law.

The experimental turn, we argue, is driven by both methodological and theoretical developments in the study of international relations and international law and, indeed, in the social sciences more broadly. In terms of methodology, the disciplines of both international relations and (more recently) international law are witnessing a shift from an emphasis on observational studies, which can be highly illuminating yet often struggle to make confident causal claims, towards an increasing use of experimental methods that seek to establish robust claims about causation through careful attention to experimental design. In this sense, the move to experiments reflects the inherent limitations of even the most technically sophisticated observational studies, and the promise of experiments is to complement (rather than replace) observational studies by testing causal claims that cannot be established through other methods.

Such a purely methodological account of the experimental turn, however, is incomplete because the shift is also part and parcel of a broader theoretical development variously known as the 'cognitive' or 'behavioural' revolution. To an increasing extent, scholars across the social sciences are calling into question the admirably parsimonious, but, in some cases, demonstrably inaccurate, rational choice models that have spread from economics to neighbouring disciplines and strongly shaped both international relations theory and the law and economics approach to international law. Drawing on advances in cognitive psychology and behavioural economics, international relations and international law scholars are developing theories and advancing policy proposals that are built on more realistic understandings of human cognition and decision making. In doing so, researchers in both disciplines have used experiments to ascertain empirically how both elites and mass publics think about and respond to international law.

While a thorough discussion of these parallel and intertwined methodological and theoretical developments is beyond the scope of this short article, we shall briefly consider each of them in turn, before turning to a careful analysis of Shereshevsky and Noah's innovative experimental study.

A *The Methodological Shift: From Observational to Experimental Studies*

In the past two decades, empirical legal studies have moved from the periphery to the centre of legal scholarship. Alongside formal, doctrinal, and normative scholarship, international law journals devote increasing space to the publication of empirical studies that attempt to understand 'the conditions under which international law is formed and has effects'.[3] Whether qualitative or quantitative, these 'observational studies' collect empirical data about phenomena such as international treaties, judicial decisions and arbitral awards in an effort to understand both the factors that shape these legal texts as well as their effects on actors in the international system. In recent years, these methods have become increasingly sophisticated and ambitious, with the

[3] Shaffer and Ginsburg, *supra* note 2.

rise of 'big data' projects capable of capturing and coding not just a sample but, rather, all of the treaties, judicial decisions or arbitral rulings in entire issue areas of international law, permitting the identification of previously unknown patterns and trends in the data.[4] This emergence of big data has in turn sparked impressive methodological innovations, such as network analysis and text as data, which have allowed empirical scholars to both describe and explain international legal phenomena. Using network methodology, for example, scholars have been able to map patterns of judicial citation and precedent, formulating and testing hypotheses about the litigation strategies of states as well as the behaviour of international courts citing legal precedent from their own and other courts' jurisprudence.[5] Similarly, using computer-driven textual analysis programs, scholars have been able to identify patterns of similarity and difference across legal texts and lines of influence across treaties and from one area of international law to another.[6]

Despite the promise of these new approaches, a growing number of scholars in both international relations and international law highlight the inherent difficulties of basing confident causal claims about international law solely on observational studies. Consider, for example, Louis Henkin's famous observation that '[a]lmost all nations observe almost all principles of international law and almost all of their obligations almost all of the time'.[7] Sceptical political scientists persuasively note that high rates of compliance with international law tell us little about international law's causal impact. High compliance rates may reflect the shallow or undemanding nature of many international agreements[8] or the fact that states self-select into international agreements, which may simply screen, rather than constrain, their signatories.[9] Going further, Adam Chilton and Dustin Tingley forcefully argue that, given the lack of variation in the 'treatment' of international law upon states, the problem of self-selection by states to legal agreements and the practical impossibility of isolating the effect of international law from other, confounding variables in the international and domestic environment, even the most sophisticated observational studies encounter

4 For an excellent introduction, see Alschner, Pauwelyn and Puig, 'The Data-Driven Future of International Economic Law', 20 *Journal of International Economic Law* (*JIEL*) (2017) 217.

5 See, e.g., Lupu and Voeten, 'Precedent in International Courts: A Network Analysis of Case Citations by the European Court of Human Rights', 42 *British Journal of Political Science* (2011) 413; Pelc, 'The Politics of Precedent in International Law: A Social Network Application', 108 *American Journal of Political Science* (*AJPS*) (2014) 547; Charlotin, 'The Place of Investment Awards and WTO Decisions in International Law: A Citation Analysis', 20 *JIEL* (2017) 279.

6 See, e.g., Allee, Elsig and Lugg, 'The Ties between the World Trade Organization and Preferential Trade Agreements: A Textual Analysis', 20 *JIEL* (2017) 333; Broude, Haftel and Thompson, 'The Trans-Pacific Partnership and Regulatory Space: A Comparison of Treaty Texts', 20 *JIEL* (2017) 391; Alschner, 'The Impact of Investment Arbitration on Investment Treaty Design: Myth versus Reality', 42 *Yale Journal of International Law* (*YJIL*) (2017) 1.

7 L. Henkin, *How Nations Behave* (2nd edn, 1979), at 47.

8 Downs, Rocke and Barsoom, 'Is the Good News about Compliance Good News about Cooperation?', 50 *International Organization* (*IO*) (1996) 379.

9 Von Stein, 'Do Treaties Constrain or Screen? Selection Bias and Treaty Compliance', 99 *American Political Science Review* (*APSR*) (2005) 611.

severe obstacles in establishing the causal impact of international law.[10] They argue that experiments can allow for a far more confident assessment of international law's causal effect, by controlling for the effects of many confounding variables that plague more complex, real-world observational studies.

In this context, a growing number of political science researchers, together with a smaller number of international legal scholars, have begun to embrace the use of experiments as a means for testing and refining theories about international law's role in international affairs.[11] Such experiments, informed by, and often modelled on, those conducted in the fields of psychology and behavioural economics, typically involve the recruitment of a subject population who are exposed, in a controlled setting, to a 'treatment', such as information about international laws or legal commitments, in order to assess the causal impact of that treatment on a dependent or outcome variable of interest. In such experiments, the subject population is divided randomly into experimental and control groups, with only the experimental group exposed to the treatment. Since the assignment of subjects to the two groups is randomized, experimenters assume that this process will generate groups that are very similar across all dimensions except for the one being studied.[12] As a result, any differences in outcomes between the groups can be assigned confidently to the treatment. Using these methods, a recent series of experiments have sought to measure the impact of international law on subjects' support for specific state policies, such as the use of force, intervention in the internal affairs of other states or the torture of terrorism suspects. In these experiments, subjects in the experimental group are provided information about international legal rules and norms, while those in the control group are not provided that

[10] Chilton and Tingley, 'Why the Study of International Law Needs Experiments', 52 *Columbia Journal of Transnational Law* (2013) 173. The authors identify five aspects of international law that make observational studies problematic, and experimental studies well suited for the purposes of testing the causal effects of international law. These are: (i) insufficient variation across states in the acceptance of many treaties or customary international national law to allow for observational testing, whereas experiments can randomly assign information about the content of international law; (ii) a short window between signing and adoption of international laws by states, which can be simulated in experiments; (iii) the existence of overlapping international and domestic (constitutional) constraints on states, which renders it difficult or impossible to isolate the impact of international law in observational studies (but not in experiments, where both treatments can be controlled; (iv) the difficulty of measuring compliance in observational studies, while such measures can be carefully designed in an experimental setting and (v) the problem of selection bias in observational studies, which can be overcome by random assignment to the 'treatment' of international law in experiments. *Ibid.*, at 181.

[11] *Ibid.* For excellent general introductions to the use of experiments in political science, see, e.g., McDermott, 'Experimental Methods in Political Science', 5 *Annual Review of Political Science* (*ARPS*) 31 (2002); J.N. Druckman *et al.* (eds), *Cambridge Handbook of Experimental Political Science* (2011).

[12] The seminal discussion of randomization in experimental design remains R. Fischer, *The Design of Experiments* (1935). To be clear, randomization renders the treatment and control groups stochastically equal (or equal on average), not identical. The law of large numbers teaches that, as the size of the groups increase, both the mean and the distribution of both groups with respect to every characteristic (save for the treatment itself) will increasingly converge. In certain contexts, researchers can use block (or stratified) randomization to ensure that treatment and control groups are balanced along important dimensions (i.e., gender, race, age) but are still randomly assigned.

information; the two groups are then compared to assess the causal impact that exposure to international law has on subjects' attitudes and behaviour.[13]

The primary virtue of experiments, then, lies in their ability to isolate the causal impact of a particular independent variable, or treatment, on a dependent or outcome variable – a characteristic known as internal validity. To the extent that the experimenter can ensure that the control and experimental groups do not differ systematically in terms of their background and demographics (thanks to the random assignment of subjects to the two groups) or in terms of their experience in the experiment (with the exception of the treatment), the experiment possesses high internal validity, and the researcher can assert with confidence that any observed difference in the dependent variable across the two groups is attributable to the independent variable or treatment.

In addition to their generally high level of internal validity, experiments have other valuable properties. Not only can experiments allow social scientists to isolate the causal effects of individual treatments on outcome variables, but they can, through repeated experiments, assess the conditions under which a particular causal relationship holds, a point we return to below. In addition, researchers can sometimes examine interactions between a given independent variable (such as international law) and another variable (such as partisanship) on a dependent variable (such as attitudes towards the use of torture),[14] although drawing valid inferences about such interaction effects often implicates complex questions of experimental design.

Of course, not every experiment is well designed, and the simple act of carrying out an experiment does not guarantee a high degree of internal validity. A large body of writings has identified a multitude of threats to internal validity.[15] For example, experimenters may inadvertently introduce bias into the experiment if and insofar as the experimenter communicates how researchers want or expect the subjects to respond, or the design of the experiment may fail to achieve balance across the treatment and control groups. Like other types of observational studies, moreover, experiments are also prone to measurement problems, and the internal validity of an experiment can be called into question by unreliable measures of experimental or

[13] Experiments along these lines are increasingly numerous. See, e.g., Chilton, 'The Influence of International Human Rights Agreements on Public Opinion: An Experimental Study', 15 *Chicago Journal of International Law* (2014) 110; Sharman *et al.*, 'Causes of Noncompliance with International Law: A Field Experiment on Anonymous Incorporation', 59 *AJPS* (2015) 146: Chilton and Versteeg, 'International law, Constitutional Law, and Public Support for Torture', 3 *Research and Politics* (2016) 1; Hafner-Burton *et al.*, 'How Activists Perceive the Utility of International Law', 78 *Journal of Politics* (2016) 167; Putnam and Shapiro, 'International Law and Voter Preferences: The Case of Foreign Human Rights Violations', 18 *Human Rights Review* (2017) 243; Wallace, 'International Law and Public Attitudes toward Torture: An Experimental Study', 67 *IO* (2013) 105; Chaudoin, 'Promises or Policies? An Experimental Analysis of International Agreements and Audience Reactions', 68 *IO* (2014) 235. See also Linos, 'Diffusion through Democracy', 55 *AJPS* (2011) 678 (testing whether knowing that an international organization recommends a particular policy increases voters' support for a proposal).

[14] See, e.g., Wallace, *supra* note 13, which identifies strong partisan differences in the effects of international law on individuals' support for the use of torture on terrorist suspects.

[15] McDermott, *supra* note 11, at 36–37 (synthesizing nine classic threats to internal validity).

outcome variables.[16] For this reason, readers of experimental studies should be cognizant of various potential threats to internal validity and assess the ways in which, and the extent to which, researchers have dealt with and minimized these threats in any particular study.

The other major limit of experiments concerns the problem of external validity – namely, the question of whether experimental results obtained in a highly controlled and simplified experiment carried out on a particular subject population are generalizable to other, more complex real-world scenarios and to the real-world subject population of interest.[17] Specifically, the problem of external validity involves at least two major and distinct questions, namely (i) the realism of experimental scenarios as proxies for real-world international relations situations and (ii) the possibility of generalizing from experimental samples to real-world international decision makers.

With respect to the realism of experiments, many students of international relations (and, even more so, international law) have been historically resistant to the use of experiments on the grounds that the great international challenges of war and peace – cooperation and conflict – cannot be replicated (or at least not ethically replicated) in experimental settings. Perhaps, for this reason, international relations scholars adopted experimental methods much later than did their political science colleagues who study domestic or comparative politics. Even for those unfazed by 'experimenting' with questions of war and peace, moreover, the challenges of ensuring external validity from experimental to real-world settings are multiple. Perhaps the most obvious problem is what we might call the level-of-analysis problem or the problem of individual versus collective decision making. The level-of-analysis problem arises from the fact that the majority of social science experiments are carried out on individuals, while state-centric theories of international law and of international relations posit states as the primary actors in the international system.[18]

In many experiments, individual subjects are instructed to imagine themselves as decision makers on behalf of states, and experimenters examine how individual subjects respond to scenarios meant to approximate interstate relations. To the extent that

[16] *Ibid.*, at 37. E.g., some experiments, including the majority of the experiments carried out by Shereshevsky and Noah, rely on self-reports from experimental subjects as measures of experimental or control variables. To the extent that these self-reports are inaccurate (e.g., if subjects are unintentionally inaccurate or deliberately lie to make themselves look better in the eye of the experimenter), the ability of the researcher to attribute causal effects to a given treatment can be compromised. See also McDermott, 'Internal and External Validity', in Druckman *et al.*, *supra* note 11, 27, at 28.

[17] External validity investigates the generalizability of results by inquiring, '[t]o what populations, settings, treatment variables, and measurement variables can this effect be generalized?' McDermott, *supra* note 16, at 35, quoting Campbell and Stanley, 'Experimental and Quasi-Experimental Designs for Research on Teaching', in N.L. Gage (ed.), *Handbook of Research on Teaching* (1963). See also Hyde, 'Experiments in International Relations: Lab, Survey, and Field', 18 *ARPS* (2015) 403, at 406–407; Hafner-Burton *et al.*, 'The Behavioral Revolution and International Relations', 71 *IO* (2017) S1, at S21–22.

[18] Even in liberal, public choice and constructivist approaches to international law, which disaggregate the state and foreground the activities and influence of interest groups, political parties, civil society and other actors, the level-of-analysis problem remains since experiments are typically conducted on individuals.

individual experimental subjects demonstrate systematic biases in decision making, such as overconfidence or a tendency to give greater weight to evidence that conforms with their prior beliefs, the results of such studies have potentially important implications for theories of state behaviour.[19] In the real world of international decision making, however, actors do not take decisions in the splendid isolation of an experimental setting but, more typically, in collective settings such as bureaucracies. As Rose McDermott notes,

> many aspects of real-world complexity are difficult to simulate in the laboratory. Cultural norms, relationships of authority, and the multitask nature of the work itself might invalidate any results that emerge from an experiment that does not, or cannot, fully incorporate these features into the environmental or manipulation. In particular, subjects may behave one way in the relative freedom of an experiment, but quite another when acting within the constrained organizational or bureaucratic environments in which they work at their political jobs. Material and professional incentives can easily override more natural psychological or professional concerns that might manifest themselves more readily in the unconstrained environment of the laboratory. Failure to mimic or incorporate those constraints into experiments, and difficulty in making these constraints realistic, might restrict the applicability of experimental results to the real world.[20]

In response to this challenge, researchers have designed experiments to simulate collective decision-making settings, 'but for skeptical readers of experimental research, the artificiality of the laboratory setting – especially when subjects pretend to be aggregate actors such as states, bureaucracies, or militaries – makes it hard to rule out the possibility that such games are not useful approximations of real-world situations'.[21]

A related concern has to do with the low stakes of experimental settings for subjects, by comparison with the enormous pressures and high-stakes consequences typical of real-world international and/or crisis decision making.[22] Recent research in behavioural economics, which posits that human beings often fall back on simple heuristics (what Daniel Kahneman has called System 1 thinking) in ordinary, low-stakes decision making while engaging in more reflective and careful weighing of options (or System 2 thinking) in real-world, high-stakes situations, provides reason to believe that the low-stakes, and potentially low-engagement, behaviour of subjects in experimental settings may differ systematically from the behaviour of actors making choices in high-pressure, high-stakes, real-world settings.[23]

This brings us, in turn, to the equally serious challenge of generalizing from commonly used samples of convenience (such as students or Internet survey subjects) to other populations, including both general publics and international legal elites, including international judges. The core challenge here is that of 'unrepresentative subject pools', which creates a disjuncture between the population to which a theory

[19] For excellent discussions of the types of biases to which international relations decision makers may be prone, see, e.g., Hyde, *supra* note 17; Hafner-Burton *et al.*, *supra* note 17.

[20] McDermott, *supra* note 11, at 39–40.

[21] Hyde, *supra* note 17, at 407.

[22] McDermott, *supra* note 11, at 39–40.

[23] Hafner-Burton *et al.*, *supra* note 17, at S22, citing D. Kahneman, *Thinking, Fast and Slow* (2015).

is meant to apply and the pool of subjects from which the experimental sample is drawn.[24] In the study of international relations and international law, theories vary in the population to which they are meant to apply. Liberal theories of international cooperation, for example, often focus on the political constraints that leaders face in making foreign policy. According to audience costs theory, for example, national elites are constrained in their foreign policy choices by mass publics who, it is theorized, impose electoral or political costs on leaders who back away from their commitments, including by violating international legal agreements.[25] Experiments using representative samples of the relevant publics provide an excellent means of testing whether exposure to knowledge about international law does, in fact, affect the foreign policy views of mass publics and, hence, whether national elites would thus feel constrained by public opinion in deciding whether to respect or violate international obligations.[26] Insofar as these studies survey representative samples of the theory's target population – namely, mass publics – they can assuage concerns about external validity.

Other theories of international law and international relations, however, take as their unit of analysis the beliefs, biases and behaviours of various groups of elites, such as foreign-policy elites of states, non-governmental organizations (NGO) or international judges. Because such elites are often difficult to recruit into experimental studies, however, researchers often rely on convenience samples, such as university students or respondents to Internet surveys.[27] In the past, a number of studies have proceeded on the assumption that experimental results from convenience samples such as university students or members of the general public are generalizable to what Emilie Hafner-Burton, Alex Hughes and David Victor call 'experienced elites'.[28]

However, to the extent that elites differ systematically from populations of convenience because of differences in background, education, socio-economic class, expertise, responsibility or experience, the ability to generalize from convenience samples to elites becomes more problematic. In response to this challenge, a growing number of scholars have explicitly addressed the problem of external validity by comparing the responses of inexperienced (for example, students) and experienced (elite) subjects. In several such studies, scholars have found that experienced elites in international relations respond to stimuli in systematically different ways from students or other convenience samples.[29] Similarly, a growing body of sophisticated experiments has

[24] McDermott, *supra* note 11, at 39.

[25] Tomz, 'Domestic Audience Costs in International Organization', 61 *IO* (2008) 821.

[26] See, e.g., the aforementioned studies by Chaudoin, Chilton, 'The Laws of War', Chilton, 'Influence', Chilton and Versteeg, Putnam and Shapiro, Wallace, all *supra* note 13. In this sense, the liberal turn towards theories emphasizing domestic public opinion and audience costs fits well with, and may help explain, the rise of survey experiments as a means of testing such theories. We thank Adam Chilton for this observation.

[27] See Hafner-Burton, Hughes and Victor, 'The Cognitive Revolution and the Political Psychology of Elite Decision Making', 11 *Perspectives on Politics* (2013) 368.

[28] *Ibid.*

[29] See, e.g., Mintz, Redd and Vedlitz, 'Can We Generalize from Student Experiments to the Real World in Political Science, Military Affairs, and International Relations?', 50 *Journal of Conflict Resolution* (2006) 757; Hafner-Burton, Hughes and Victor, *supra* note 27.

found that law students and domestic judges differ systematically in their ability to apply legally relevant norms in politically charged legal disputes.[30] These findings are sobering and suggest caution, at a minimum, in assuming that findings from students and other convenience samples can necessarily 'travel' to international elites, including international judges, in particular.

In light of these concerns, scholars whose theories address the attitudes or behaviour of international elites have increasingly sought to recruit subjects from groups of political, military, economic, foreign policy, NGO, and judicial elites, assigning subsets of these groups at random to experimental and control groups, in an effort to increase the external validity of experiments.[31] More generally, while sceptics about experiments sometimes depict external validity problems as a fatal challenge to experimental studies of international relations and international law, defenders argue that problems of external validity can be addressed, most notably by replicating and extending experiments to incorporate more realistic scenarios, more reliable measures and more and different types of subjects in order to determine whether and how results from experimental studies remain robust when replicated in new ways and with more realistic subject populations.[32] The correct response to concerns about external validity, advocates suggest, is not despair but, rather, replication.[33]

For these reasons, proponents of experimental methods warn against rejecting experimental methods out of hand.[34] Indeed, rather than understanding themselves as facing a dichotomous choice between experiments and other scholarly methods, researchers can integrate experimental methods into broad, multi-method research programs, employing experiments deliberately where they can compensate for the deficiencies of other methods. 'Experimental methodology proves most useful', according to McDermott, 'when investigators desire the ability to draw clear causal inferences, seek to resolve discrepancies in findings reported using alternative methodologies, or wish to uncover underlying micro-foundational psychological processes in particular'.[35]

[30] See, e.g., Kahan *et al*., '"Ideology" or "Situation Sense"? An Experimental Investigation of Motivated Reasoning and Professional Judgment', 164 *University of Pennsylvania Law Review* (2016) 349 (experiment finding that judges of diverse cultural outlooks converge on legal outcomes in politically charged cases but that law students polarize along the same lines that divide legally untrained members of the public); Redding & Reppucci, 'Effects of Lawyers' Socio-Political Attitudes on Their Judgments of Social Science in Legal Decision Making', 23 *Law and Human Behavior* (1999) 31 (experiment finding that pre-existing opinions and political outlooks on the death penalty do not influence judges' decisions regarding admissibility of social science studies on deterrent effect of the death penalty, but does influence law students' assessments of admissibility).

[31] Hyde, *supra* note 17, at 407–409.

[32] For a good general discussion of replication as a response to problems of external validity, see McDermott, *supra* note 16, at 37–38.

[33] *Ibid.*, at 34: 'External validity results primarily from *replication* of particular experiments across diverse populations and different settings, using a variety of methods and measures' (emphasis in original). For a sophisticated analysis of the replication literature across disciplines, see Firth, Hoffman and Wilkinson-Ryan, 'Law and Psychology Grows Up, Goes Online, and Replicates', available at http://scholarship.law.upenn.edu/faculty_scholarship/1884.

[34] See, e.g., Hyde, *supra* note 17, at 404.

[35] McDermott, 'New Directions for Experimental Work in International Relations', 55 *International Studies Quarterly* (2011) 503, at 511.

McDermott's final point about 'microfoundations' is particularly important for scholars of international law and international relations. Researchers in both fields may be inclined to reject experiments primarily because of a level-of-analysis problem since both the empirical phenomena studied by such scholars and many of the theories that they employ take place at the macro level, focusing on collective actors such as states (rather than individuals) and on large systems or structures (rather than agents). Yet, as Susan Hyde has argued, scholars can plausibly link macro-level systemic theories and phenomena to micro-level individual responses in carefully designed experiments. Macro-level phenomena, such as the outbreak of war or the impact of international institutions on international cooperation, 'frequently take place on a smaller scale' through the decisions and actions of individuals, which can be studied empirically.[36] Furthermore, many macro-level theories have 'implicit or explicit micro-level implications', which can be made explicit and tested through experimental methods.[37] In this way, experiments may not only speak to the 'big questions' of international relations and international law but also contribute to the theoretical development, as well as the testing, of systemic research programs.[38]

In doing so, moreover, experimental researchers can make use of at least three different types of experiments: laboratory experiments, survey experiments and field experiments.[39] All three of these types share the central feature of attempting to assess the causal impact of a particular independent variable (or treatment) on a dependent variable (or outcome) through random assignment to experimental or control groups, but each of the three is carried out in a distinct setting, and each embodies a different set of trade-offs among desiderata such as internal and external validity.

Laboratory experiments arguably maximize internal validity by allowing investigators to randomly assign subjects to the experimental and control groups as well as various potentially confounding variables that can be held constant. Laboratory experiments are also economical and allow researchers to study interactions between subjects, facilitating the testing of game-theoretic models of cooperation and conflict that are common in international relations and international law scholarship. However, because of the need to bring subjects physically into the laboratory, such

[36] Hyde, 'The Future of Field Experiments in International Relations', 628 *Annals of American Academy of Political and Social Science* (2010) 72, at 74.

[37] *Ibid.*

[38] *Ibid.* For a contrasting perspective, see Mearsheimer and Walt, 'Leaving Theory Behind: Why Simplistic Hypothesis Testing Is Bad for International Relations', 19 *European Journal of International Relations* (2013) 427 (increasing emphasis on empirical hypothesis testing leads to less attention to theory, which in turn impoverishes the discipline).

[39] For excellent discussions of these three types of experiments, including their respective strengths and weaknesses and their contribution to international relations and international law scholarship, see, e.g., Hyde, *supra* note 17; Chilton and Tingley, *supra* note 10, at 219–236. All of these are distinct from so-called 'natural experiments' in which the treatment variable is not assigned by the researcher but is nevertheless distributed in a way that approximates randomization and is thus conducive to quasi-experimental analysis. See, e.g., T. Dunning, *Natural Experiments in the Social Sciences: A Design-Based Approach* (2012); Dunning, 'Improving Causal Inference: Strengths and Limitations of Natural Experiments', 61 *Political Research Quarterly* (2008) 282.

experiments are often administered to 'convenience samples' such as university students, raising concerns about external validity and generalization to other populations such as international political and legal elites.[40]

Survey experiments, by contrast, can reach more widespread and potentially representative subject populations, by embedding randomized experiments within public opinion surveys. Like traditional surveys, survey experiments expose respondents to a series of prompts and questions and record and code their responses for analysis. In addition, however, the embedded experiment assigns members of the survey sample at random to an experimental group that is exposed to some written or visual treatment or to a control group that does not receive the treatment, and the investigators then record the two groups' responses to subsequent questions. Technological advances and the increasing use of Internet surveys have decreased the cost of survey experiments dramatically, and survey instruments are increasingly used to test hypotheses about the ways in which exposure to particular prompts (including information about the content of international law) influences the opinions and decisions of survey respondents.[41] Survey experiments allow investigators to reach a wider pool of potential subjects, including elite subjects, which in principle facilitates external validity, although the ability to generalize from survey experiments to real-world situations remains an open question.[42]

Field experiments, finally, feature researchers intervening directly in real-world contexts, randomly introducing a 'treatment' to experimental groups, while withholding that treatment from a control group, to determine whether the treatment has an effect on outcomes 'on the ground'. First developed and applied in American politics, particularly to measure the impact of various get-out-the-vote procedures on actual voter turnout, field experiments have more recently spread to comparative and international politics. For example, Susan Hyde worked with the Carter Center to conduct a field experiment to determine the impact of international election observers on voting outcomes in the 2004 Indonesian presidential election. Specifically, Hyde allocated the distribution of election observers among polling districts at random, finding that voting outcomes differed systematically across the two groups.[43] Other international relations studies, often undertaken in cooperation with national or international organizations such as the World Bank or USAID, have sought to determine the efficacy of international development programs by allocating such programs at random among regions within a country and comparing outcomes between the control and treatment groups.[44] In a recent, innovative set of field experiments that

[40] Hyde, *supra* note 17; Chilton and Tingley, *supra* note 10, at 222–226.

[41] See, e.g., Chaudoin, Chilton, 'The Laws of War', Chilton, 'Influence', Chilton and Versteeg, Putnam and Shapiro, all *supra* note 13.

[42] See, e.g., Barabas and Jerit, 'Are Survey Experiments Externally Valid?', 104 *APSR* (2010) 226; Findley *et al.*, 'External Validity in Parallel Global Field and Survey Experiments on Anonymous Incorporation', 79 *Journal of Politics* (2017) 856.

[43] Hyde, 'Experimenting in Democracy Promotion: International Observers and the 2004 Presidential Elections in Indonesia', 8 *Perspectives on Politics* (2010) 511.

[44] See the excellent discussion in Hyde, *supra* note 36, at 76.

tested the effects of international law on the behaviour of non-state actors, Michael Findley and his colleagues contacted (via email) hundreds of incorporation services, allocating at random a series of six different messages, including one that specifically mentioned the international legal ban on anonymous incorporation. The authors found, dishearteningly, that specifically raising international law had no effect on the willingness of firms to provide the service in question.[45]

Properly conducted field experiments tend to possess relatively high external validity while, in principle, confronting many of the same internal validity problems that face survey experiments. Despite their advantages, however, in practice field experiments are relatively difficult to conceptualize, frequently raise difficult logistical (and sometimes ethical) issues and often require substantial resources to conduct properly. Nevertheless, as Chilton and Tingley suggest, '[w]ell-designed field experiments directly test existing theories in a way that has the potential to generate evidence convincing to both scholars and policy-makers',[46] and their use in both international relations and international law seems likely to grow.

B *The Theoretical Shift: From Rational Choice to Behaviourism*

Thus far, we have presented the experimental turn as a methodological development, driven, in part, by the limitations of observational methods and, in part, by the promise of experimental methods in establishing causation. The shift to experiments, however, is also associated with a broader theoretical development, the so-called cognitive or behavioural revolution, which has spread from the discipline of psychology into the disciplines of economics, political science and, most recently, law. In all three of these disciplines, rational-choice approaches, derived from neo-classical economics,[47] had come to dominate not only neo-classical economic thought but also the dominant strands of rationalist international relations theory[48] as well as the law and economics

[45] Findley, Nielson and Sharman, 'Causes of Noncompliance with International Law: A Field Experiment on Anonymous Incorporation', 59 *AJPS* (2015) 146; Findley, Nielson, and Sharman, 'Using Field Experiments in International Relations: A Randomized Study of Anonymous Incorporation', 67 *IO* (2013) 657. More recently, Katerina Linos and Tom Pegram published a quasi-experiment in which they examined how 107 states responded to differences in the language used in different provisions of a soft law UN General Assembly resolution calling for the creation of national human rights institutions. They found that variations in agreement language can have substantial effects on state behaviour. Specifically, they found that both democratic and authoritarian states followed 'firm terms' closely but that authoritarian states either ignored or reduced their efforts on flexibly specified tasks. Linos and Pegram, 'The Language of Compromise in International Agreements', 70 *IO* (2016) 587.

[46] Chilton and Tingley, *supra* note 10, at 235.

[47] See, e.g., Gary Becker's canonical definition of rationality: '[A]ll human behavior can be viewed as involving participants who [1] maximize their utility [2] from a stable set of preferences and [3] accumulate an optimal amount of information and other inputs in a variety of markets.' G.S. Becker, *The Economic Approach to Human Behavior* (1976), at 14, quoted in van Aaken, 'Behavioral International Law and Economics', 55 *Harvard International Law Journal* (2014) 421.

[48] See, e.g., the institutionalist and regime theoretic analyses in S. Krasner, *International Regimes* (1982) and R.O. Keohane, *After Hegemony* (1984), imported into international law by Abbott, 'Modern International Relations Theory: A Prospectus for International Lawyers', 14 *YJIL* (1989) 335.

movement in the legal academy.[49] In both international law and international relations, scholars theorized states (and, in some cases, non-state actors such as individuals, firms, NGOs and international organizations) as egoistic rational actors seeking to maximize their utility in strategic interactions in a condition of anarchy.[50] Drawing in large part on game-theoretic models, these rationalist scholars formulated and tested rational-choice theories that posited international institutions and law as solutions to prisoner's dilemma and collective action problems, making cooperation possible by establishing rules and norms and providing information about the state of the world and the behaviour of other players.[51]

Beginning in the 1970s, however, a behavioural revolution took place in the economic profession from which rationalist analyses had earlier spread.[52] Put simply, a growing number of economists concluded that the classical assumptions of egoistic rationality in *homo economicus* were incompatible with a growing body of empirical evidence emerging, in particular, from the study of cognitive psychology. For cognitive psychologists, including most famously Daniel Kahneman and Amos Tversky, and for the behavioural economists whom they influenced, the classical assumptions of rationality, including the stable and transitive nature of actor preferences, the systematic updating of beliefs and the process of decision making, were to be subject to systematic empirical testing, largely through laboratory experiments.[53] Several decades later, cognitive psychologists and behavioural economists have discovered multiple systematic, pervasive departures in human behaviour from the ideal type of economic rationality, which, in turn, prompted the development of new and alternative micro foundations for the study of a wide range of human behaviour.

While a complete survey of these departures from rational-choice models is beyond the scope of this short article, Hafner-Burton and her colleagues usefully distinguish three sets of what they call 'nonstandard' preferences, beliefs and decision-making procedures uncovered in several decades of behavioural research.[54] With respect to

[49] On the application of the rationalist law and economics approach to domestic and international law, respectively, see, e.g., R. Posner, *Economic Analysis of Law* (2nd edn, 1977); Dunoff and Trachtman, 'Economic Analysis of International Law', 24 *YJIL* (1999) 1.

[50] As Hafner-Burton and her coauthors summarize such models, '[i]ndividuals are assumed to maximize expected utility by determining the payoffs attached to all possible outcomes, assessing their probabilities, updating information on those probabilities, and choosing the strategy with the highest expected return. In game-theoretic models, equilibrium outcomes are generated out of the choices of two or more players in the game'. Hafner-Burton *et al*., *supra* note 17, at S6.

[51] See, e.g., A. Guzman, *How International Law Works: A Rational Choice Theory* (2008); J.P. Trachtman, *The Economic Structure of International Law* (2008).

[52] For an account by a central actor in this story, see R.H. Thaler, *Misbehaving: The Making of Behavioral Economics* (2015).

[53] Kahneman and Tversky famously identified a series of biases and heuristics present in human decision making under uncertainty, which they analysed in a 1979 paper under the rubric of 'prospect theory'. See Kahneman and Tversky, 'Prospect Theory: An Analysis of Decision under Risk', 47 *Econometrica* (1979) 263. For an influential, updated and popularized account, see Daniel Kahneman, *Thinking, Fast and Slow* (2011).

[54] Hafner-Burton *et al*, *supra* note 17, at S7–S13. An alternative categorization of the biases discovered by behavioural economists distinguishes what Jolls and her colleagues call bounded rationality, bounded willpower and bounded self-interest. See, e.g., Jolls, Sunstein and Thaler, 'A Behavioral Approach to Law and Economics', 50 *Stanford Law Review* (1998) 1471, as updated and applied to international law in van Aaken, *supra* note 47, at 426–435.

preferences, experimental research identified systematic deviations from the textbook model of rationally ordered preferences, including in attitudes towards risk (that is, being risk averse with respect to gains but risk acceptant with respect to losses); 'hyperbolic discounting' of future costs and benefits and, perhaps most intriguingly, the existence of 'social' (altruistic or socio-tropic) preferences, which led individuals in public goods experiments to be more cooperative than rationalist game-theoretic models, with their assumption of pure self-interest, had predicted.[55]

With respect to beliefs, rational choice models had assumed 'that beliefs are formed and updated in a way that avoids systematic error'.[56] By contrast, psychologists and behavioural economists identified widespread and systematic misperceptions in human behaviour, including the overweighting of available evidence, failure to seek out or pay attention to other evidence and chronic overconfidence. As Kahneman and Tversky had demonstrated early on, human beings turned out to be poor intuitive statisticians, prone to relying on systematic biases and flawed heuristics that frequently generate suboptimal choices.[57]

Finally, with respect to decision-making processes, investigators discovered systematic deviations from the predictions of expected-utility theory in both individual and collective decision making. Individually, experimental subjects were shown to be subject to both 'cold' cognitive shortcuts and heuristics and 'hot' emotional influences and, thereby, limited in their ability to calculate expected costs and benefits of possible actions. For example, multiple studies have shown that the framing of a choice (for example, in terms of gains or losses or opt-in versus opt-out clauses) can induce subjects to make dramatically different choices even where the underlying payoffs are identical. Similarly, collective decision making was also shown to be subject to processes, such as the well-documented phenomenon of 'groupthink', which augmented individual tendencies towards selective attention to information and overconfidence.[58]

From its pioneering application in behavioural economics, this new behavioural or cognitive approach to human behaviour spread rapidly to other disciplines.[59] In the political science subfield of international relations, scholars influenced by the behavioural revolution in psychology and economics drew attention to systematic deviations from perfect rationality in areas as diverse as international security and crisis decision making (where miscalculation, selective attention to evidence and overconfidence were widespread) to international political economy (where individual preferences on issues such as trade were shown to be influenced by considerations such as race

[55] Hafner-Burton *et al.*, *supra* note 17, at S9–S10.

[56] *Ibid.*, at S11.

[57] *Ibid.*, at S11–S12.

[58] *Ibid.*, at S12–S17

[59] *Ibid.*, at S2: '[A] new behavioral revolution has swept across the social sciences in the last few decades. With origins in psychology, of course, psychological models have fueled the dramatic growth of behavioral economics and are now gaining traction in political science as well. The defining characteristic of this revolution has been the use of empirical research on preferences, beliefs, and decision-making to modify choice- and game-theoretical models.'

and gender, which were far removed from models of open economy politics).[60] Indeed, much of the international relations experimental scholarship already reviewed was motivated by an attempt to study the impacts of 'non-standard' preferences, beliefs and decision-making processes on the behaviour of actors in international politics. Applications of behavioural insights have been widespread in international relations, including in the areas of international security (focusing, in particular, on crisis behaviour), international political economy (focusing in large part on actors' trade and economic policy preferences) and, most recently, international law (seeking to determine the impact of international law on actors' choices and behaviour).[61] While earlier waves of international relations scholarship had, of course, focused on the measurement and explanation of political behaviour, '[w]hat is new in today's behavioral revolution is the explosion of experimental research',[62] which has blossomed in international relations scholarship alongside prospect theory and other cognitive approaches. Behavioural international relations scholars have not only followed the theoretical lead of their predecessors in psychology and behavioural economics but also sought to test their theories using similar experimental methods and designs.

By the late 1990s, the behavioural revolution had reached the discipline of (domestic) law, with a call by Christine Jolls, Cass Sunstein and Richard Thaler for a new 'behavioral law and economics' that would 'explore the implications of actual (not hypothesized) human behavior for the law'.[63] Building on the insights of cognitive psychologists and behavioural economists, advocates of this new school argued that a behavioural approach held the promise of describing more accurately the way that '[h]umans' – as opposed to '[e]cons' – actually interacted with the law. The aim of such an approach, they point out, is not to jettison the theoretical framework of law and economics but, rather, to strengthen it by basing economic models on more realistic micro foundations that reflect how people really behave. Such models, in turn, hold the prescriptive promise of using public law and policy to create a 'choice architecture' that would 'debias' human behaviour and 'nudge' individuals in a non-coercive fashion towards individually and socially desirable ends.[64]

More recently, a small, but growing, body of international law scholars has started to explore the implications of behavioural discoveries for the design and working of

[60] The literature on cognitive psychology, prospect theory and international relations is large and ever-growing. See, e.g., McDermott, 'Prospect Theory in Political Science: Gains and Losses from the First Decade', 25 *Political Psychology* (2004) 289; Mintz, 'Behavioral IR as a Subfield of International Relations', 9 *International Studies Review* (2007) 152; Goldgeier and Tetlock, 'Psychological Approaches', in C. Reus-Smit and D. Snidal (eds), *Oxford Handbook of International Relations* (2008) 462; Hafner-Burton, Hughes and Victor, *supra* note 28; Hafner-Burton *et al.*, *supra* note 17. For an excellent analysis of how individuals' preferences towards international trade diverge from the predictions of traditional, rationalist models, see Alexandra Guisinger, *American Opinion on Trade: Preferences without Politics* (2017).

[61] Hafner-Burton *et al. supra* note 17.

[62] *Ibid.*, at S2.

[63] For useful introductions, see Jolls, Sunstein and Thaler, *supra* note 54, at 1476; C. Sunstein (ed.), *Behavioral Law and Economics* (2000).

[64] Thaler and Sunstein, *Nudge: Improving Decisions about Health, Wealth, and Happiness* (2009); Jolls and Sunstein, 'Debiasing through Law', 35 *Journal of Legal Studies* (2006) 199.

international law.[65] While acknowledging the challenges of applying the insights of individual psychology to the workings of states and other corporate actors such as courts, advocates of 'behavioral international law'[66] or 'behavioral international law and economics'[67] suggest that awareness of individual cognitive biases, heuristics and other 'psychological kinks in rationality'[68] can help to better explain how and why states and individuals consent to international law;[69] how they design international treaties;[70] whether and how they comply with their international legal commitments;[71] how international judges interpret the law[72] and a multitude of other research questions. This body of scholarship remains in its infancy, and its authors have only begun to engage in systematic experimental research,[73] but the existence of

[65] See, e.g., Woods, 'A Behavioral Approach to Human Rights', 51 *Harvard International Law Journal* (2010) 51; Galbraith, 'Treaty Options: Towards a Behavioral Understanding of Treaty Design', 53 *Virginia Journal of International Law* (2013) 309; Poulsen and Aisbett, 'When the Claim Hits: Bilateral Investment Treaties and Bounded Rational Learning', 65 *World Politics* (2013) 273; Poulsen, 'Bounded Rationality and the Diffusion of Modern Investment Treaties', 58 *International Studies Quarterly* (2013) 1; van Aaken, *supra* note 48; Broude, 'Behavioral International Law', 163 *University of Pennsylvania Law Review* (2015) 1099.

[66] Broude, *supra* note 65.

[67] Van Aaken, *supra* note 47.

[68] Broude, *supra* note 65.

[69] See, e.g., Galbraith, *supra* note 65 (arguing that state decisions to accept the jurisdiction of the International Court of Justice [ICJ] are influenced by framing effects from treaties); Poulsen and Aisbett, *supra* note 65 (arguing that state consent to bilateral investment treaties with investor–state dispute settlement provisions can be explained by states' systematic underestimation of the probability of being subject to litigation – an error not corrected until the first 'claim hits').

[70] Van Aaken, *supra* note 47, at 456–470 (reviewing and comparing the implications of rationalist contract theory and behavioural approaches to treaty design).

[71] *Ibid.*, at 471–480 (exploring the implications of behavioural findings regarding the purported effectiveness of retaliation, reciprocity and reputation as drivers of state compliance); Broude, *supra* note 65, at 36–41 (exploring the effects of uncertainty and framing effects on military decisions under international humanitarian law).

[72] Broude, *supra* note 65, at 32–36 (exploring the possibility of panel or conformity effects in international judicial decision making).

[73] Existing works often consist largely of literature reviews and broad research agendas (Brooks, Broude and van Aaken) or of behavioural theories tested using observational data (Galbraith, Poulsen and Poulsen and Aisbett). Galbraith, for example, argues that the framing of 'treaty choice' in an area such as the acceptance of ICJ jurisdiction can vary, from implicit opt-outs (in the case of reservations), to explicit opt-outs, to explicit requirements to opt in. Drawing on behavioural insights about the status quo bias, she predicts – and finds in a large-n observational study of state choices – that states are far more likely to accept the compulsory jurisdiction of the ICJ with respect to treaties in which such acceptance is the default condition, as opposed to treaties in which they must explicitly opt in to such jurisdiction. This is a well-designed study. At the same time, it illustrates the limitations of observational studies discussed above. Specifically, some readers will wonder whether Galbraith's highly suggestive findings are due to selection effects. It may be that states negotiate opt-in clauses in more sensitive areas, in which case the much lower acceptance of ICJ jurisdiction may be endogenous to the nature of the issue that the treaty addresses rather than resulting from the framing effect of the treaty itself. See Galbraith, *supra* note 65, at 314, 342 (acknowledging the concern but dismissing it as implausible). This is a good example of a question for which experimental studies might allow a researcher to isolate the effect of the 'treatment' of treaty framing (opt-out versus opt-in provisions), while holding the nature of the issue constant and thus allowing for more confident causal claims, supplementing Galbraith's rigorous observational studies.

a behavioural strand of international legal theory is likely to further drive the development of experimental studies in the years to come.

2 Evaluating Shereshevsky and Noah

Judges and scholars have engaged in long-standing doctrinal and normative debates over when, if ever, treaty interpreters should have 'recourse … to supplementary means of interpretation, including the preparatory work of the treaty'.[74] Shereshevsky and Noah's intervention into this debate seeks to test empirically whether exposure to preparatory work impacts treaty interpretation, 'even under a rule that prohibits its use'.[75] To do so, the authors expose experimental groups of law students and 'international law experts' to the 'treatment' of preparatory materials while withholding those materials from control groups. Within each experiment, the authors distinguish between and compare responses of subjects who self-report that use of preparatory materials for interpretative purposes is legally permissible and those who report that it is not permissible.[76] Shereshevsky and Noah interpret their results to suggest that exposure to preparatory work makes a measurable and statistically significant impact on treaty interpretation by law students, even when they believe they are not permitted to take the preparatory work into account. For the 'international law experts', however, the data suggests an ability not to allow preparatory work to influence the interpretation of legal text in circumstances where the rules forbid the consideration of preparatory work for interpretative purposes.

How convincing are these findings and how relevant are they to the real world of international legal interpretation? To explore these two questions, we briefly analyse Shereshevsky and Noah's study with an eye to the two primary criteria – internal and external validity – by which researchers generally assess experimental scholarship.

A *Internal Validity*

Internal validity, once again, refers to the ability to draw causal inferences from a study – namely, that an independent variable causally impacts a dependent variable. In the case of Shereshevsky and Noah's study, this involves an assessment of whether, as the authors claim, they are able to isolate the causal impact of their independent variable (exposure to preparatory materials) and of an interaction variable (attitudes about the permissibility of using such materials) on a dependent or outcome variable (judgment about a state violation of an international treaty), while holding all other potentially confounding variables constant. To determine whether Shereshevsky and Noah meet this standard, let us restate briefly their basic research design.

[74] Vienna Convention on the Law of Treaties (VCLT) 1969, 1155 UNTS 331, Art. 32.
[75] Shereshevsky and Noah, *supra* note 1, at 1291.
[76] *Ibid.*, at 1296, n. 33; 1297, n. 36.

Put simply, Shereshevsky and Noah run a series of experiments on two distinct subject samples: Experiments 1a and 1b on law students, Experiment 2 on legal elites. Within each of these samples, subjects are randomly assigned to an experimental or a control group: subjects in the experimental group are exposed to preparatory materials, while those in the control group are not. Afterwards, each subject is asked to interpret the treaty in question, and the authors then compare the two groups to determine whether exposure to preparatory materials affected the subjects' treaty interpretations. Shereshevsky and Noah, however, are not only interested in whether exposure to preparatory materials affect treaty interpretation, but they also wish to know whether preparatory materials influence legal interpreters 'even under a rule that prohibits its use'. They therefore collect data on an additional variable – namely, whether the subject in question believes that the use of preparatory materials is allowed, or not allowed, when interpreting the treaty in question, and they look to see whether the effect of exposure to preparatory materials is moderated by subjects' views about the legal permissibility of resorting to such materials. Indeed, throughout the article, Shereshevsky and Noah present the results of their experiments separately for the 'allowed' and 'not allowed' groups – the former in Tables 1, 3, 5 and 7 and the latter in Tables 2, 4, 6 and 8.

Among the student populations in Experiments 1a and 1b, Shereshevsky and Noah find that those who are exposed to preparatory materials are substantially less likely to find a treaty violation than those who are not so exposed. This effect, moreover, holds for those who believe that preparatory work is not permitted (see Table 2) as well as among those who believe that it is permitted, suggesting that, like jurors unable to ignore inadmissible evidence, law students are influenced by preparatory materials even when they believe that they should not be.[77] These findings are intriguing, but some readers may wonder how useful it is to compare results between one group of individuals who were exposed to preparatory materials and another group that was not. As Shereshevsky and Noah correctly note, it is generally accepted that parties to international disputes 'will always refer the tribunal to the *travaux*'.[78] For this reason, the question Shereshevsky and Noah test for possesses limited real-world purchase. The more relevant question is whether, among those who have been exposed to preparatory materials, those who think they are permitted to use these materials for interpretive purposes read the treaty differently than those who think that they are not permitted to use preparatory materials.

[77] For ease of exposition, we focus our discussion on Experiment 1a, although the same basic findings hold for Experiment 1b, also on students. By contrast, the authors find that in Experiment 2, legal elites behave quite differently. The 'allowed' group is, once again, less likely to find a violation when exposed to preparatory materials, but the 'not allowed' group appears not to be influenced by exposure, suggesting that legal elites are more able to ignore such information than are students. We return to this point under the rubric of external validity below.

[78] Shereshevsky and Noah, *supra* note 1, at 1289–1290 (quoting Aust).

Shereshevsky and Noah appear to be addressing this question in their discussion of Hypothesis 3, which posits that 'the participants' views regarding their ability to use preparatory work under the VCLT [Vienna Convention on the Law of Treaties] rules of interpretation can moderate the effect of the exposure to preparatory work, even if they do not fully eliminate the effect'.[79] To test this claim, the authors compare the decisions of those who think the use of preparatory materials is 'allowed' to those who think it is 'not allowed', following exposure to the treatment of preparatory materials. Comparing the top row of Table 1 to the top row of Table 2, for example, they find that, among those exposed to preparatory work, those who believe it could be used in interpretation find no violation more frequently (80.7 per cent) than those who believe that preparatory materials cannot be used (56.6 per cent). This would seem to suggest that preparatory work has a greater impact on those who believe that its use is allowed, although the authors report that, '[n]evertheless, a logistic regression on participants' decisions did not reveal a significant interaction between exposure to preparatory work and their position on resorting to preparatory work under the specific scenario'.[80] It is therefore unclear whether the subjects' beliefs about the permissibility of using preparatory materials does or does not influence their use of those materials in treaty interpretation.

However, note that in this comparison the authors are looking only at the subjects in the experimental group, namely those who were exposed to preparatory materials. Looking further into these data, however, reveals a puzzling anomaly. If we consider the subjects in the control group who were not exposed to preparatory materials, roughly 55 per cent of these subjects who thought resort to preparatory work was permissible found no violation;[81] in striking contrast, only about 9 per cent of the individuals who thought resort to preparatory work was not permitted believed there was no violation.[82] We are not told whether the dramatic difference between the two groups is statistically significant, but the result (and similar differences between the 'allowed' and 'not allowed' groups who were not exposed to preparatory materials in other experiments) raises a puzzling question: why would an individual's attitude towards the use of preparatory materials impact treaty interpretation when the individual was not exposed to these materials?

This unexplained finding, in turn, raises concerns regarding the internal validity of Shereshevsky and Noah's experiments. Specifically, it raises the possibility that something about the beliefs regarding preparatory work predisposes subjects to interpret treaties differently independent of the effect of exposure to preparatory materials. Recall that, in most of the authors' experiments, each subject self-selects into the

79 *Ibid.*, at 1294, VCLT, *supra* note 74.
80 Shereshevsky and Noah, *supra* note 1, at 1300.
81 *Ibid.*, at Table 1.
82 *Ibid.*, at Table 2.

'allowed' and 'not allowed' groups by answering explicit questions on this point.[83] Shereshevsky and Noah treat these responses, not implausibly, as measures of a discrete trait – namely, the subject's view specifically of the permissibility of using preparatory materials to inform treaty interpretation. Yet, there are two reasons to question whether the design of the study, and the measurement of this specific variable, actually allows the authors to isolate the influence of this particular interpretive stance on each subject's decisions about state violations of treaty law. First, each subject was asked her views about the permissibility of considering preparatory materials after issuing her decisions. It is, therefore, possible that the subjects' answers to this interpretive question were influenced by, and perhaps attempts to justify post hoc, their prior decisions on whether a violation had occurred.

Second, and equally troubling, it is also possible that each subject's self-reported views on the permissibility of considering preparatory materials reflect, and are endogenous to, their broader approaches to legal interpretation. Indeed, this possibility – that the subjects' self-reported attitudes about preparatory materials are capturing broader interpretive differences – might explain the otherwise puzzling finding that these differences correlate with different treaty interpretations even when the subjects are not exposed to such materials. To be sure, these concerns do not negate the authors' findings about the impacts of preparatory materials on treaty interpretation. But they do suggest that further studies are necessary before one can confidently conclude that Shereshevsky and Noah's experiments have truly isolated the impact of belief in the permissibility of using preparatory materials on treaty interpretation – that is, whether the experiments possess high internal validity.

B *External Validity*

By contrast with internal validity, which refers to the design and conduct of the experiment itself, external validity refers to whether the results obtained in experimental settings can be generalized to more complex real-world situations and to the broader, real-world populations of interest. To their credit, Shereshevsky and Noah acknowledge a number of challenges to the external validity of their study.[84] We applaud the authors' candour in this regard and would simply highlight three external validity challenges that Shereshevsky and Noah's studies share with other experimental studies of

[83] Specifically, each subject was asked, after deciding and justifying her decision about whether the law was violated, a series of questions about judicial interpretation. *Ibid.*, at 1296. The 'allowed' group, in this context, included 'participants who held to the corrective approach to interpretation and those who held to the traditional approach but determined that in the specific scenario it is possible to use the preparatory work to determine the meaning under the conditions of the articles of the VCLT' (at 1296, n. 33). The 'not allowed' group, by contrast, included 'participants who reported that in cases of a clear and reasonable interpretive result following the application of article 31, it was impermissible to use preparatory work to determine the meaning of the text' (at 1297, n. 36). The exception to this practice of using self-reported attitudes towards the permissibility of preparatory work is in Experiment 1b, in which the control and experimental groups were primed with different versions of the VCLT rules, allowing or not allowing the use of preparatory materials in legal interpretation (at 1300–1302).

[84] *Ibid.*, at 1311–1312.

international law and that suggest some caution in interpreting Shereshevsky and Noah's results pending further replication studies.

First, as the authors note, survey experiments typically present subjects with a hypothetical 'scenario' or 'vignette' depicting an abstracted version of a real-world problem and then expose subjects to an international law 'treatment', such as information that a particular course of action is forbidden by international law, before asking the subject to express an opinion or a decision about the scenario in question. By necessity, both the vignettes and the international law stimuli are simplified and abstracted, raising questions about whether they are useful proxies for the more complex situations and legal stimuli to which real-world decision makers are exposed.[85] In this respect, we commend Shereshevsky and Noah for using vignettes that are closely patterned on real-world cases as well as interpretative 'rules' that are closely patterned on the VCLT. Nevertheless, as the authors acknowledge, the cases were deliberately selected, and the vignettes deliberately constructed, to maximize the contrast between the apparent textual meaning of treaty provisions and the meaning implied by preparatory materials. In most real-world cases, the gap between treaty text and preparatory work is unlikely to be so large. Using vignettes with stark contrasts is clearly justifiable in a pioneering study of a new research question, yet it raises the question whether real-world adjudicators might behave differently (perhaps ignoring preparatory material more effectively) where the divergence between it and the treaty text is not as dramatic as that found in Shereshevsky and Noah's experiments. One way to eliminate this potential threat to external validity is to run experiments using cases and vignettes for which the gap is smaller.

Second, Shereshevsky and Noah encounter another ubiquitous external validity challenge, namely generalizing from individual responses in experiments to the collective decision making that typically characterizes decisions by organizational and collective actors like states and courts. All of Shereshevsky and Noah's experimental subjects made decisions on their own, whereas international tribunals typically consist of multi-member panels that engage in collective deliberations and decision making. Substantial evidence suggests that the process of group deliberations impacts decisions and, more specifically, that 'group decision-making has been found to moderate the influence of inadmissible evidence',[86] making the question of collective decision making one obvious frontier for future experimental research into questions of treaty interpretation.

Third, and perhaps most significantly, is the question of whether it is possible to generalize from experimental subjects – and, especially, subjects of convenience such as students – to real-world decision makers like international judges. Shereshevsky and Noah claim that 'conducting experiments on law students as proxies for legal experts is an accepted practice in empirical legal research'.[87] Given the practical difficulties of

[85] E.g., vignettes necessarily create much less immersion in a case than actual judicial decision making and do not require written reasons supporting a result, a requirement that can both constrain judges and promote accountability in the resolution of real world disputes.

[86] *Ibid.*, at 1312.

[87] *Ibid.*, at 1302.

recruiting international judges, we appreciate why analysts would use student subjects in experiments. Nonetheless, we believe that generalizing from undergraduate students to international judges, particularly on a complex activity such as legal reasoning, is deeply problematic. As discussed above, an increasing number of experiments suggest that university students and experienced elites differ systematically in their decision-making styles and outcomes[88] and, specifically, that judges and law students differ systematically in their ability to apply legal rules, including on the admissibility of evidence, in light of their prior political and cultural commitments. These experiments suggest that the differences between judges and law students are particularly salient for an activity like legal reasoning, which relies upon deep contextual knowledge gained over years of professional practice and which can be honed and improved as one gains experience over time.[89]

Indeed, the article's central finding – that the ability not to use preparatory materials when relevant rules preclude its use differs between law students and legal experts – suggests precisely that experienced elites behave very differently than law students when engaging in legal interpretation. This result is both encouraging and sobering. It is substantively encouraging in that it suggests that international legal elites (who are presumably a better proxy for international judges) appear able to ignore preparatory materials if they believe that the relevant legal rules provide that consideration of those materials is impermissible. It is methodologically sobering, however, with respect to studies that use samples of convenience, such as students, as proxies for experienced legal elites, such as international judges, because Shereshevsky and Noah's findings provide further reason to believe that the results of experiments on students are unlikely to 'travel' to real-world legal elites. We therefore read Shereshevsky and Noah's fascinating study as a cautionary tale about the external validity of experiments using students as proxies for legal elites and as an example of 'best practice' in which investigators carefully test their hypotheses not only on samples of convenience but also on more realistic elite samples.

3 Conclusion

Both the cognitive or behavioural revolution and its accompanying use of experimental studies have in the past half-decade reached the study of international law by way of psychology, economics and political science. Advocates of experimental research argue persuasively that well-executed experiments possess a dual promise. From a theoretical perspective, they promise to build a new study of social phenomena (including international law) not on the basis of neo-classical economic assumptions of actor rationality but, rather, on the basis of empirically documented regularities in human behaviour. In the bluntest possible terms, they make possible the construction

[88] E.g., Hafner-Burton, Hughes and Victor, *supra* note 27.

[89] Karl Llewellyn famously explained how judges and lawyers develop a 'situation sense', or a perceptive facility, formed through professional training and experience, permitting them to assimilate disputes to recurring 'situation-types' that indicate their proper disposition. E.g., K. Llewellyn, *The Case Law System in America*, edited by P. Gewirtz, translated by M. Ansaldi (1989).

of theories and the offering of pragmatic proposals for reform based not on how individuals should interact with international law but, instead, on how they do interact with it. From a methodological perspective, experiments offer the promise of rigorously testing the causal impact of international law on human behaviour in ways that are impossible for observational studies of a complex empirical world to achieve. For both of these reasons, the use of experimental techniques in international law scholarship holds real promise and is only likely to grow with time.

At the same time, conceiving and executing well-designed experiments on international law, such as that by Shereshevsky and Noah in these pages, is extremely difficult. In this sense, experimental research in international law today stands roughly where large-n empirical legal studies stood several decades ago, requiring both experimental researchers and consumers to approach experimental studies with a sophisticated understanding of their promise and limits. Experimental studies in international law need to be analysed with an eye to both internal validity – to ensure that the experiment is truly isolating the effect of specific independent variables – and external validity – to maximize the prospect that the results of a controlled experiment 'travel' to real-world actors and contexts. We have discussed these challenges, both in the abstract and as applied to Shereshevsky and Noah's pioneering study of the use of preparatory materials in international legal interpretation. We hope that their article, and our reply, will encourage a broader discussion of the uses and limits of experiments on international law.