LEGAL CONSTRUCT VALIDATION: EXPANDING EMPIRICAL LEGAL SCHOLARSHIP TO UNOBSERVABLE CONCEPTS

David S. Goldman[*]

[*] Adjunct-Professor, Loyola University Chicago College of Law; Staff law clerk, United States Court of Appeals for the Seventh Circuit. (David_Goldman@ca7.uscourts.gov).

INTRODUCTION

A consensus has been growing in recent years that empirical legal scholarship is an important tool for informing policy decisions.[1] Empirical study of legal concepts allows debate on fundamental questions that is informed by the actual impact of law on behavior rather than conjecture and an appeal to "common sense." But despite the recognition and broader acceptance of the benefits of empirical legal scholarship, many of the most fundamental concepts of law such as incentives, deterrence, and even justice, are not directly measurable and, therefore, have been overlooked in empirical study. Fortunately, legal scholars can adopt the notion of construct validation---the way psychologists overcome the difficulty in measuring unobservable psychological phenomenon, called constructs---to develop empirical tests that measure these unobserved legal models and expand the current horizons of quantitative study of legal concepts. Using the meta-theories of construct validation, researchers can infer the effects of unobserved constructs and can thus modify foundational legal theories based on empirical evidence rather than speculation. I propose a standardized procedure based on current understandings of construct validation that can be used to measure intangible legal constructs: 1) develop generalized legal theories, 2) infer hypotheses from those theories, 3) design experiments to test the hypotheses, and 4) modify the general theories based on the research results.

As a first step, researchers must develop theories about the effects of law that are specific enough to provide direction for empirical testing, while still taking a generalized perspective not confined to specific laws. Generalized theories are necessary to allow multiple inferences to be drawn that elucidate more detail of the unobserved latent variable being explored. For instance, rather than developing a theory only about the impact of the First Amendment, a more generalized theory will address the influences of all constitutional rights. A generalized theory may state that populations tend to exhibit more freedom when specific rights are embedded within that nation's constitution. Empirical studies of freedom of speech can then provide particularized support for the broader constitutional theory, and can also interact with studies of other rights.

---

[1] *See, e.g.*, Tracey E. George, An Empirical Study of Empirical Legal Scholarship: The Top Law Schools, 81 IND. L.J. 141 (2006) ("Empirical legal scholarship (ELS) is arguably the next big thing in legal intellectual thought."); American Association of Law Schools Annual Meeting: Empirical Scholarship--What Should We Study and How Should We Study It? (Jan 3-7, 2006), available at http://www.aals.org/am2006/index.html; Lee Epstein & Gary King, The Rules of Inference, 69 U. CHI. L. REV. 1, 2 (2002); Richard H. McAdams & Thomas S. Ulen, Introduction: Symposium: Empirical and Experimental Methods in Law, 2002 U. Ill. L. Rev. 791.

Once a generalized theory has been devised, researchers can then infer sets of hypotheses to test measurable aspects of those theories. For example, based on a theory that constitutional rights lead to more freedom, one might hypothesize that if a nation has a constitutional freedom of speech then the media may produce more commentary disapproving of the government. While the actual *freedom* that the press feels is impossible to measure, the proportion of articles disapproving of government actions is relatively easily to calculate. But a useful hypothesis suggests more than just an empirical test; it also should have the potential to undermine the theory. The necessity for critical hypotheses is borne from the current philosophical understanding that theories can never be fully verified. Hence, while empirical evidence that sustains a theory can reaffirm the theory's legitimacy, a critical hypothesis can serve one of two purposes: if supported, the hypothesis can undermine an incorrect theory, but if proven empirically inaccurate, that same hypothesis can help to rule out a potential a criticism. By eliminating potential criticisms, a researcher can more forcefully demonstrate the validity of the overarching theory.[2]

Next, the research must devise experiments that can test the accuracy of the hypotheses. Ideally, experiments would use randomized subjects for testing so that the results were likely caused by the treatment being studied rather than a bias in the selection process. For instance, if a researcher wanted to study the effects on prison terms of the United States Sentencing Guidelines, the researcher would most like to randomly assign some convicted felons to be sentenced under the guidelines and an equivalent group to receive sentencing without using the guidelines. But it is often not feasible to randomize subjects on empirical legal studies for practical, if not necessarily constitutional, reasons; most legal researchers, therefore, use various forms of quasi-experimentation that allow testing when subjects can not be randomized. Quasi-experimental designs allow inferences to be made about causal relationships when randomization is not feasible. The researcher wishing to study prison terms, for example, could use a "time-series design" to analyze sentences that were dolled out both before and after the Sentencing Guidelines went into effect.[3] But because

---

[2] Moreover, a researcher should not test only one hypothesis because the results from that single test can be influenced by a questionable supporting theory rather than the one targeted by the study. Each theory is supported by numerous auxiliary theories; for instance, many current economic theories of law are based on Ronald Coase's famous theory of transactions, which in turn, relies in part on a rational-actor theory of behavior. Hence, a test of a single hypothesis concentrated on a prevailing theory of law and economics may uncover negative results, not because of a problem with the theory being examined, but because of problems with the rational-actor model. But if only one hypothesis is tested, this possibility may not be recognized.

[3] Donald T. Campbell, et al., *Quasi-Experimental Designs*, *in* Methodology and

the experiment is not truly randomized, the researcher needs to be careful to identify plausible threats to internal validity (such as the possibility that other changes in the law during that same time period affected sentences).[4] These threats can be controlled by adjusting design elements or statistical techniques to account for the threats.[5] Finally, causal inferences can be made when the results form a coherent pattern, either within the study or as compared to outside knowledge.[6]

Once the various hypotheses have been tested, the overarching theory should be modified to reflect the new knowledge gained through the empirical data. The theory may have accurately predicted all the results of the various experiments, in which case the theory is supported but no new information is provided to the field. But if the theory's predictions are not completely accurate, the theory should be modified to incorporate this new information. Theories are not falsified every time disconfirming data is discovered, nor do fields go through seismic paradigm shifts where old theoretical frameworks are thrown out in favor of an entirely new system. Rather, through an iterative process of testing and retesting, theories about unobservable legal constructs can be modified based on inferences from the results of experimentation.

It is important to note, however, that empirical quantitative analysis cannot supplant qualitative reasoning; while empirical results can help inform legal policy-making, it is still necessary to debate and understand theories on a qualitative level. For example, empirical testing could indicate that torture could lead to improved intelligence gathering, but society must still decide whether it can morally justify such tactics. Similarly, evidence of the deterrent effect of capital punishment does not by itself substantiate the penalty.

The rest of this paper is organized as follows. Part II examines the development of philosophies of science to provide background of the current understandings of theory development. Part III then describes the history of psychometrics to demonstrate how psychology has applied philosophies of science to overcome some of the same issues that empirical legal scholarship is now facing. Finally, part IV details my proposed approach for incorporating the most relevant aspects of this related field into legal study.

Epistemology for Social Science191, 201 (1988).

[4] William R. Shadish, *The Empirical Program of Quasi-Experimentation*, *in* Research Design: Donald Campbell's Legacy 13, 16 (Leonard Bickman ed. 2000).

[5] *Id*. at 17.

[6] *Id*. at 18-19.

I. PHILOSOPHY OF SCIENCE

Before moving on to describe a system for construct validation for legal studies, it is first necessary to understand the philosophical underpinnings on which those concepts are based.

A. Sir Karl Popper's theory of verification and falsification

A common starting point when addressing the philosophy of science is the philosopher Karl Popper's idea of falsification; a theory is scientific if it can be tested and has the potential to be proven wrong.[7] The central philosophical problem that concerned Popper was the difficulty of demarcation: how is science distinguishable from pseudo-science? Popper concluded that a true scientific theory should be corroborated through exposure to empirical tests that have the potential of disproving the theory, and a theory is better corroborated when it has survived more difficult tests.[8]

Popper believed in an asymmetry between the ability to verify a theory and the ability to prove a theory false; specifically, a scientific theory can never be proven completely correct, but evidence that undermines a theory can conclusively prove that theory wrong.[9] Even a well-established theory can be proven wrong at any time if contrary evidence is discovered. For instance, we may believe that every time an object is dropped it will fall towards the Earth, but even this theory could be falsified if the next item we drop fell up. As a result, a hypothesis that can provide support for a theory is never as useful as one that can undermine the theory because the former can never prove a theory true, but the latter can definitively establish that it is false. Scientists should concentrate, then, on testing negative hypotheses because each time a negative hypothesis fails, one more potential criticism has been ruled out and the theory has been further corroborated.

But critics of this philosophy have pointed out that if this philosophy is correct, then history should be laden with critical studies that have falsified the predominant theory of the time.[10] But, in fact, most important theories were falsified countless times before they were discarded.[11] Under a falsificationist view, some of the most important theories in history were

---

[7] Karl Popper, The Logic of Scientific Discovery (Karl Popper et al. trans., 2002).
[8] Paul E. Meehl, *Theoretical Risks and Tabular Astricks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology*, *in* Selected Philosophical and Methodological Papers 1, 18 (1991).
[9] Popper, *supra* note 7, at 18-19.
[10] Brendan Larvor, *Lakatos: An Introduction* 50 (1998).
[11] *Id*.

not actually scientific. Proponents of Popper's philosophy, then, are forced to distort history and arbitrarily label certain falsifying experiments as "crucial," while ignoring others.[12]

Popper's idea that a theory can be completely discredited by any negative evidence has been labeled a type of justificationism.[13] Justificationists believe that a scientific theory can be fully proven by empirical evidence.[14] "In brief, [justificationism] is the view that the way to criticize an idea is to see whether and how it can be justified."[15] Justificationsists believe that certain facts exist and if scientists can establish that a particular theory does not reflect those facts, then the theory should be discarded.[16] For example, the premise of an argument may be that all observable planets travel in an elliptical orbit, and the conclusion is that all planets must travel in an elliptical orbit.[17] For a justifcationist, this is an invalid theory because it overstates the known facts. But for a nonjustifcationist, the conclusion stated above may still have value to science as an unjustified hypothesis rather than as a justified statement about nature. By reversing the syllogism, the conclusion can be a stated as a hypothesis, and the premises are data supporting the hypothesis. Now, contrary evidence will not mean that the theory needs to be rejected outright, but may need to be conditionally rejected or revised to reflect the new data.[18] This revision can proceed indefinitely as new information is collected. Philosophy has generally moved away from the justificationist theories, and philosophers now endorse various versions of nonjustifactionism.[19]

B. Thomas Kuhn's Theory of Paradigm Shifts

One such nonjustificationist theory is historian Thomas Kuhn's theory of scientific paradigms. After researching the history of scientific developments, Kuhn noted a distinction between mature and immature sciences. In immature sciences, many competing schools of thought exist, but no single idea dominates the field. Because an accepted theory has not been adopted, scientists operating in immature fields base their research on

---

[12] *Id*. at 42.
[13] William Warren Bartley, III, The Retreat to Commitment 104-05 (2d ed. 1984).
[14] Herbert Feigel & Albert Blumberg, *Logical Positivism: A New European Movement*, J. of Philosophy 28, 281-96 (1931).
[15] Bartley, *supra* note 13, at 186.
[16] *Id*., at 186-87.
[17] *Id*. at 191.
[18] *Id*. at 196.
[19] Greg Smith, *On Construct Validity: Issues of Method and Measurement*, 17 Psychological Assessment 396 (2005); Bartley, *supra*, note 13, at 194.

many different schools of thought. But in exchange for this freedom to choose their own theories, the scientists must give up the benefits from relying on established research supporting the broad theory. Because no accepted idea for a field exists, the scientists must reassert every basic concept on which any new research relies. For example, before a single concept for gravity was accepted, any publication of new research had to start with a thorough explanation of the scientist's particular perspective on gravity and how those assumptions could affect the results. Hence, an immature field moves along essentially at random because no controlling force exists to direct future exploration. (This state of an immature field also seems to describe the current state of legal academics. It is often necessary to spend a significant portion of a law review article stating the assumptions and background on which a new theory is based because, without it, a reader will not know what school of thought and definitions the author is employing).

But researchers in an immature science do not need to despair that their field will flounder with no direction indefinitely; a science can mature, not necessarily because of a new field-shattering discovery, but from a theoretical proposal that is influential enough to persuade members of opposing schools of thought. Once a controlling theory is accepted, mature fields are then guided and driven by this single idea. For a new proposal to bring about maturity in a field, then, the new idea must also be open-ended enough so that future scientists have room to explore concepts controlled by the idea. Kuhn labeled these driving forces "paradigms," which specifically are "a set of recurrent and quasi-standard illustrations of various theories in their conceptual, observational, and instrumental applications."[20] Essentially, a paradigm is an accepted worldview or framework under which scientists explore their field. In mature sciences, paradigms do not evolve slowly, but rather are subject to seismic transitions in thought, what Kuhn called paradigm shifts.[21] When scientists go about their day-to-day work in a mature field they usually work within the dominant paradigm and engage in "normal science," which is "research firmly based upon one or more past scientific achievements."[22] Most scientists in a field generally subscribe to the same general paradigm, and perform experiments to define the details of the paradigm. As the scientists discover new information they adjust the prevailing theories to accommodate their results.

---

[20] Thomas S. Kuhn, The Structure of Scientific Revolutions 43 (3d ed. 1996).

[21] *Id*. at 12-13. Prior to Kuhn, the word "paradigm" had been used only in connection with grammar and linguistics. Since Kuhn's adoption of the word to science, however, the broader use of the word has become more prominently accepted.

[22] *Id*. at 10.

1. Normal science and the maturation of a field

Kuhn described the development of electrical theory as a typical example of the maturation of a scientific field. Until the mid-eighteenth century, the study of electricity floundered without a dominant paradigm.[23] Before that time scientists did not agree on a unifying concept for electricity and so each new experiment had to stand on its own accord without leaning on previous research for support.[24] As is common in other immature fields, early researchers worked in one of many different schools of thought that described some of the known physical properties of electricity, such as the belief that electricity consisted of several different fluids, but no single view could completely explain all empirical observations.[25] Without a dominant paradigm, facts could not be prioritized, evaluated, or criticized and physicists did not have a basis with which to devise future experiments. Development of the field proceeded with no direction and therefore discoveries were made slowly and essentially at random.[26] But in the mid-eighteenth century, Benjamin Franklin proposed that electricity may be a single fluid consisting of positive and negative charges. Franklin's theory could explain more characteristics of electricity than any other theory of the time, and it therefore grew to become widely accepted among other researchers. It also suggested new avenues for future exploration and could therefore guide scientists to conduct experiments that would lead to more useful results.[27]

Kuhn suggests that Franklin's fluid-theory for electricity was the first dominant paradigm in the field.[28] Once the paradigm was established, physicists no longer had to waste time reexamining the basics of electricity, and could begin to narrow their experiments to systematically explore the details of the paradigm. Now that the field had matured, the scientists could engage in the "normal science" of the time. Some of the discoveries by eighteenth-century physicists could have been made only by operating within this paradigm.

Kuhn postulated that each new fact uncovered through normal science that supports the prevailing paradigm makes it more likely that theory is correct. "Normal science does not aim at novelties of fact or

---

[23] *Id*. at 12-13. Other examples of paradigmatic works are Aristotle's *Physica*, Ptolemy's *Almagest*, Newton's *Principia* and *Opticks*, Lavoiser's *Chemistry*, and *Lyell's Geology*.
[24] *Id*. at 13.
[25] Franklin theorized that electricity was a fluid that operated differently under different pressures. He said that electrical fluid under one level of pressure was "positive" and under another was "negative."
[26] Kuhn, *supra* note 20, at 16.
[27] *Id*.
[28] *Id*. at 18.

theory and, when successful, finds none."[29] As scientists continue to engage in normal science, however, they may encounter anomalous results that cannot be explained using the theories of the current paradigm. The positive results are like a weight on a scale in favor of the existing paradigm, and the negative results act like counterweights against the existing paradigm. Kuhn argued that once an anomaly is uncovered, scientists explore the anomaly. As more becomes known about the odd event, researchers modify the existing theories to account for the new information. Eventually, the anomaly will become a predictable part of the current paradigm.[30]

In this way, Kuhn's conception of normal science is in stark contrast to Popper's philosophy of falsification in which a negative result could individually undermine an entire theory. According to Popper, scientists constantly engage in critical experiments that test the very foundations of their belief structure; their basic purpose was to disprove the underlying assumptions and embrace contradictory evidence. To the contrary, Kuhn believed that scientists almost never ask fundamental questions, but merely accept the existing paradigm as a foundation for further research. Kuhn provided the following example of an anomalous result being incorporated into the existing paradigm: A scientist notices that while performing an experiment on cathode rays, a screen in his lab has an unexpected glow; further investigations of the anomalous glow then leads to the discovery of an unknown form of radiation called x-rays that were not predicted by the current paradigm.[31] This scientist as well as other researchers then perform research on the x-rays and adjust the existing theories to reflect the new information. Scientists can now predict the occurrence and effects of x-rays. What was once an anomalous result, is now a predictable part of the dominant paradigm.

2. Paradigm shifts

But Kuhn believed that when enough anomalous results accumulate, they can reach a critical mass that tips the balance of commonly accepted theories away from the predominant view and throws the field into a state of crisis. When a field is in crisis, it no longer operates under a common set of beliefs, but instead resembles the immature state where scientists constantly need to restate their basic assumptions to communicate a new idea.[32] In a time of crisis, scientists will begin retesting assumptions that had been

[29] *Id*. at 52.
[30] *Id*. at 52-53.
[31] *Id*. at 57.
[32] *Id*. at 72.

widely-accepted, and a new competing paradigm may slowly develop. Eventually, the believers in the new paradigm will challenge the hold-outs from the older system and the better system will prevail. A better scientific paradigm is one that accounts for more of the known data and more accurately predicts future outcomes. For example, in the late nineteenth century Newtonian physics and concepts of space failed to account for increasingly accurate measures of movement of celestial bodies. Astronomy's inability to accurately predict celestial movement had pushed the field into a state of crisis. This crisis only subsided with the publication of Albert Einstein's special theory of relativity. When the new standard prevailed, a "scientific revolution" or "paradigm shift" had taken place; Einstein's relativity theory was a paradigm shift away from Newtonian physics.

Kuhn's philosophy has drawn its own criticisms, however. First, his ideas are not useful to predict future paradigm shifts.[33] One cannot say ex ante whether science is on the verge of a new paradigm shift. And if a revolution does appear imminent, no one within the field is able to tell how much more information is necessary to complete the transition. The difficulty in prediction arises because Kuhn's idea is based on historical observations and can only be applied in an ex post position.

Other critics have attacked Kuhn's idea that two paradigms in the same field must be incommensurate, meaning scientists cannot hold to two different worldviews at the same time.[34] Kuhn believed that physicists cannot simultaneously believe that Newton's and Einstein's theories are both correct. The incommensurate component of Kuhn's theory also implies that it is not possible to prove whether the field is better off for undergoing a revolution; the new paradigm may carry with it an entirely new definition for the science.[35] Kuhn did not believe that there is a "true" scientific theory; merely that one theory may be a better predictor of future outcomes. This concept has been given the disparaging label of "relativism."[36] Moreover, scientific discoveries do not actually develop in well-defined leaps. Instead, they evolve slowly and at any given time many scientists may believe aspects of both old and new theories. As even Kuhn

---

[33] Thomas S. Ulen, *A Nobel Prize in Legal Science: Theory, Empirical Work, and the Scientific Method in the Study of Law*, 2002 U. Ill. L. Rev. 875, 885 (2002).

[34] *See, e.g.,* Donald Davidson, *On the Very Idea of a Conceptual Scheme*, 47 *Proceedings and Addresses of the American Philosophical Association*, 5-20 (1973 - 1974) (attacking the idea of conceptual relativism---the idea that reality is dependent on the framework with which it is viewed. Davidson argues that supporters of a certain worldview can always discuss other worldviews, and thus it is incoherent to claim that two views are completely incommensurate).

[35] Larvor, *supra* note 10, at 42.

[36] *Id*. at 43.

noted, scientists do not generally believe that they are developing a new paradigm, but usually think they are simply building on established ideas.

C. Current Understandings in Philosophies of Science

Philosophers have come to understand that theories do not stand by themselves, but actually rely on many supporting, or auxiliary theories. Any theory about the interaction between celestial objects, for example, is based on theories about gravity, inertia, mass, the make-up of the space around the objects, etc. The success of any given theory therefore depends on the truth of its auxiliary theories.[37] Therefore, a negative result from an experiment may indicate that a theory is false, but it could also mean that an auxiliary theory is incorrect. Because a researcher confronting a negative result can never be certain whether the result reflects the theory being examined or a supporting concept, any particular theory is never fully confirmed.

A key implication of this understanding is the importance of criticism in theory development.[38] Because a theory can never be fully proven correct, criticism may be more useful than additional support. A theory that describes all known events of a given phenomenon may not necessarily be correct; other examples that the theory could not describe may not yet have been discovered or a different theory may describe the events even better. If scientists were to look only for more positive examples that are described by the theory, they would supply little new information for the field. For example, Newton's theories accurately described the data known at his time, but as more accurate observations were collected the theory's flaws became apparent. If instead of looking for supporting evidence, scientists attack a position critically they can rule out potential criticisms and choose which of competing theories best describe empirical evidence.

Legal academics have developed a similar method of continuous critique and modification of theory. What sets the legal academy apart from the sciences, however, is that the sciences have embraced the power of empirical tests to support debates about theories. In contrast, law relies on criticism of theory in a more abstract sense. This difference between law and other academic fields results, at least in part, on the difficulty of measuring legal concepts. Physicists can use a ruler to measure the distance an object travels, anthropologists have developed sophisticated tests to measure the age of relics that they uncover, and chemists can use scales to weigh the product of a chemical reaction. But lawyers do not have a tool

---

[37] Smith, *supra* note 19, at 397-98.
[38] W.B. Weimer, *Notes on the Methodology of Scientific Research*, 40 (1979).

that can measure civil liberties, security, or freedom. Lawyers seeking to empirically test existing theory face the difficulty of figuring out what to measure.

Fortunately, the complexities of measuring abstract concepts are not unique to law, and are common in the social sciences. Psychology in particular has developed sophisticated methods with which to measure elusive notions such as when is a person actually depressed as opposed to just sad. This subfield of psychology, called psychometrics, has developed over the last century to help researchers deal with the difficulty of treating concepts that are not easily quantifiable in a scientific manner.

II. PSYCHOMETRICS

Towards the end of the nineteenth century the scientific world was still reacting to Darwin's publication of *On the Origin of Species by Means of Natural Selection, or The Preservation of Favoured Races in the Struggle for Life* (commonly known as *The Origin of Species*). Today, his theory of evolution proposed in the book is well-known. But less well-known is that Darwin also concluded, based on this theory, that humans consisted of two subgroups---savages and civilized---and the distinction between the two was based at least in part on intelligence. Although Darwin could conclude that some people must have advanced to a higher intellect than others, up to that point no one could empirically test that conclusion.[39] The challenge of actually measuring a person's intelligence intrigued many philosophers. The difficulty arose because raw intelligence was an abstract trait of a human mind and an obvious metric did not exist. The philosophers eventually began testing other measurable traits that they believed would have some correlation to intelligence; even if it was not possible to directly measure how smart a person was, it was possible to estimate based on a set of related characteristics. The first attempt at a test measured visual and auditory acuity and several other psychophysical variables.[40] Although the specific metrics used in the first intelligence test turned out not to be very accurate, the idea of using measurable traits to estimate an immeasurable quality formed the basis of modern psychometrics.

Researchers expounded on this early work by attempting to find characteristics with higher correlations to known indicators for intelligence,

---

[39] Unfortunately, Darwin's conclusion that that intelligence was a genetic trait also led to misuse of the intelligence tests. The eugenics movement was based on the belief that the quality of the human race could be improved by selective breeding based on the results of intelligence tests.

[40] John Rust & Susan Golombok, *Modern Psychometrics*, 5 (1999).

such as to students' grades.[41] By the end of the first decade of the twentieth century a consensus was achieved with regard to several testing techniques for measuring intelligence, and a test developed by Alfred Binet was practically applied for the first time to identify slower students in Paris schools for assignment in "special" classes. Binet assembled a variety of testing criteria into a single test that was easy to administer, and teachers confirmed the accuracy of the test based on their own personal evaluations of their students. The test was so effective that derivatives were used for over sixty years to help identify the mentally retarded.[42]

A. The Introduction of Construct Validation

Since the 1950s, psychologists have been developing similar but more sophisticated tests with which to measure other unobservable phenomena, such as intelligence, depression, happiness, and other mental characteristics. One of the key difficulties in assessing these traits, however, is determining whether the tests are in fact measuring the correct feature. In 1955, psychologists Lee Cronbach and Paul Meehl articulated the necessity of construct validity;[43] validity is the degree to which a given test actually measures the trait, or construct, that it is intended to measure. They proscribed that "[c]onstruct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured."[44] In other words, when faced with an abstract concept with no convenient metric, an investigator must first find a measurable criterion that can be used to approximate the concept. But it is not enough to just find a single standard that may approximate the trait: it must also be valid, which means it should closely correlate with the construct being studied.[45]

Because of the inherent difficulty in determining whether a given tangible criteria accurately reflects the value of the trait it is being used to

---

[41] Pearson developed the Pearson Product-Moment Correlation Coefficient for analyzing the data.

[42] In the United States, a version of the test called the Stanford-Binet was widely used. But because psychologists struggled for an adequate definition for intelligence, experts could not come to a consensus on what characteristics the tests were actually measuring. A. Anatatasi, Psychological Testing, 5th ed. Macmillan, 67 (1982). By the 1950s, many psychologists believed that the tests' scores did not indicate anything more than the ability to perform well on the tests themselves.

[43] Lee J. Cronbach & Paul E. Meehl, *Construct Validity in Psychological Tests*, 52 Psychological Bulletin (1955), *reprinted in* Paul E. Meehl, Psyhcodiagnosis: Selected Papers 3 (1973). Meehl actually considered himself a neo-Popperian.

[44] *Id.* at 5.

[45] *Id*. at 3.

estimate, Cronbach and Meehl suggested that researchers "bootstrap" their way from a single criterion to achieve increasingly valid tests.[46] Once a valid criterion is found, it can be used to develop more valid tests. Essentially, a test based on a valid criterion may become more accurate than even the original criterion itself.[47] For example, people first recognized the physical quality of temperature because certain objects felt different than others, and they used the sense of touch to measure this characteristic. But early observers did not yet fathom that this feeling could be accurately gauged or scientifically tested beyond simple feel. Researchers eventually noticed, however, that when certain elements felt hot, they expand and when they felt cool they contracted. Further, certain elements, such as mercury, went through significant noticeable expansion and contraction within the temperature range in which people lived. And the expansion correlated well with the original test for temperature---feel. Most important, though, was that the expansion correlated with not just one person's judgment of temperature but with anyone's perspective. It turned out that mercury was an even more accurate gauge of certain temperature-related events such as boiling and melting points than touch ever was. In the end, scientists were able to use feel as a measurement for temperature to bootstrap a more valid test using mercury. Similarly, Binet's intelligence tests were accepted because their results correlated well with teachers' expectations about their students' intelligence. In time, I.Q. tests were viewed as a more valid measurement for intelligence than teachers' expectations.[48] The validity of more sophisticated I.Q. tests could later be verified by comparing them to Binet's test. But as these examples demonstrate, a test---like a scientific theory---can never be proven completely valid, and establishing validity must be an ongoing process.

It became apparent that like Popper's philosophy of falsification, a vital component of construct validation is review and criticism.[49] Without critical assessment, a measurement's validity can never be confirmed because it would not be compared to other known indicators. And when a metric cannot be tested through observable events, supporters may be tempted to rationalize its effectiveness. Cronbach and Meehl warned however that "[r]ationalization is not construct validation;"[50] researchers must develop hypotheses that critically test the measurement and a single hypothesis alone is not sufficient.

---

[46] *Id*. at 11.
[47] *Id*..
[48] *Id*.
[49] *Id*. at 17-18.
[50] *Id*. at 18.

B. Modern Understandings of Construct Validation

Using a method called the multi-trait multi-method approach, ("MTMM") a researcher can further validate a metric by not only observing factors that should have positive correlations with the test, but also investigate factors that should be unrelated to the trait if the theory is correct.[51] By using this approach, the validity of a test can be determined by measuring several traits---some that should have high correlations with the construct being studied and some that should not--- and by using several different methods. This way a researcher can ensure that the results are an accurate reflection of the trait being studied and not a reflection of a related trait or an artifact from a particular experimental method.[52]

To illustrate, consider a researcher who wants to assess a theory regarding mathematical reasoning and requires the use of a new test for math skills.[53] To determine whether the new test is correctly measuring the proper traits, the researcher hypothesizes that the test is valid if the results correlate well with older, accepted tests. But even though the researcher may be correct that a high correlation can indicate validity, this relationship alone is not definitive. Without more information it is impossible to determine whether the test is truly measuring only math skills, or if an unrecognized third factor, such as reading skill, is actually influencing both the score on the old test and the new. Hence the researcher must develop and test additional hypotheses regarding the possible influence that reading or other skills may have on the new test. If further observations reveal that the test's results have a high correlation with previous math scores, but an even higher correlation with previous reading scores, then the test is probably not a valid measure for mathematical reasoning alone. Rather, the test may actually be measuring reading skills, and those people who are better able to read the problems may score higher than those with better math abilities simply due to a better understanding of the test. But without critically exploring this alternate hypothesis, the researcher would have based future research on a potentially invalid test.

If, after testing, the researcher discovers that the hypothesis is not supported, a reexamination of the overarching theory is necessary. But this examination is not simple; scientific theories are often based on many auxiliary theories.[54] Results can also be affected by elements of the test

---

[51] Smith, *supra* note 19, at 396; Rust & Golombok, *supra* note 40, at 73.
[52] Rust & Golombok, *supra* note 40, at 73.
[53] This example is derived from examples presented by Greg Smith. *See* Smith, *supra* note 19; *see also* Rust & Golombok, *supra* note 40.
[54] Meehl, *supra* note 8, at 20.

itself.[55] Hence, when empirical observations do not conform to theoretical predictions, the negative results do not necessarily mean that the theory is wrong.[56] Instead, it may be that the theory is correct and the test is not accurate. On the other hand, it is possible that the test is accurate and the theory is not completely correct, that both the theory and the tests are not adequate, or that the auxiliary theories are wrong. Unfortunately, because of these ambiguities, even if the hypothesis is supported, the researcher still cannot be completely confident in the theory.[57] Once again, it is necessary to test the supporting theories in addition to testing the theory of interest.[58]

When an experiment does confirm a hypothesis, the positive results can help establish the test's validity, but Cronbach and Meehl originally claimed that the test cannot last in the face of negative results.[59] In this way, their conception of construct validity resembles Popper's theory of verification.[60] But just as philosophers moved away from Popper's justificationist view, psychologists have realized that tests are not fully invalidated in the face of negative information. Rather, measurements must undergo constant revision.[61]

To validate a test of a construct, psychologists can follow a five-step method.[62] First, a theoretical construct, such as intelligence or happiness, should be specified. Next, an informative hypothesis should be developed that will add to the knowledge of the field rather than reaffirming existing ideas. Then research experiments can be designed that critically test the hypothesis and the data's correlation with the hypothesis's predictions can be calculated. Finally, the theory should be revised to reflect the new knowledge gained from testing the hypothesis. Under this method, theories constantly evolve to reflect the updated knowledge of the field.

III. MEASURING UNOBSERVABLE LEGAL CONSTRUCTS

Sciences that involve abstract concepts (in particular, the social sciences) share the common obstacle of measurement that is now confronting those wishing to approach legal questions scientifically. But because this dilemma is common among many different fields, legal scholars are able to use the experiences from the other fields to overcome these issues. This interdisciplinary approach is based on the idea of

---

[55] *Id*.
[56] *Id*.
[57] Smith, *supra* note 19, at 396.
[58] Meehl, *supra* note 8, at 21.
[59] Cronbach & Meehl, *supra* note 43, at 18.
[60] *See* discussion *supra* Part II.A.
[61] Smith, *supra* note 19, at 396.
[62] *Id*.

consilience, which is the "'jumping together' of knowledge by the linking of facts and fact-based theory across disciplines to create a common groundwork of explanation."[63] It is the idea that when one studies all levels of scientific phenomena, similar principles can be discerned; distinct fields of study may not be as discrepant as one might expect. For example, one of the most prevalent of these similarities is that many fields decide among multiple possible explanations for a given outcomes by using Occam's razor, which states the principle of parsimony; essentially, parsimony refers to a preference for the least complex explanation for a given outcome. Because of these underlying symmetries between fields (such as Occam's razor) it is possible to solve some of the most complex problems in one field by integrating the knowledge from other disciplines.

Accordingly, I propose extending the concept of construct validation from psychology to law. Construct validity looks for convergence and divergence of operations to make inferences about latent variables based on measurable traits that are influenced by the construct. For example, while happiness is unobservable, it is possible to measure happiness by measuring behaviors such as the number of times people smile or the frequency of complimentary statements. Similarly, while deterrence may be an unobservable legal construct, it is possible to infer that a law has had a deterrent effect by measuring the frequency of unwanted behaviors or the number of arrests. As mentioned in the introduction, however, it is important to remember that empirical observations should not be the sole basis for policy decisions. Even if a supported theory predicts a certain outcome, lawmakers must still decide whether those laws comport with the societies qualitative values.

The method I propose primarily involves four steps: first, a researcher must develop generalized theories about the influence of law on society. Generalized theories should predict behavioral outcomes that one would expect as the result of a given type of law. Because unobservable legal constructs can only be empirically studied by noting divergent and convergent operations, a generalized theory should suggest as many different operations as possible; as more operations are tested, confidence in the presence of the latent variable should increase. Second, based on the theory, the research must infer hypotheses that can challenge and provide useful information about the theory and specify operations to test hypotheses. Although many legal empirical studies have devised and tested hypotheses, some do not, and even fewer test rival hypotheses or address possible weaknesses in an accepted theory.[64] But, as explained above, experiments that simply reaffirm a hypothesis are much less informative

[63] Edward O. Wilson, Consilience: The Unity of Knowledge 8 (1998).
[64] Epstein & King, *supra* note 1, at 9, 37, 76-80.

than those that address criticism of the theory.[65] Third, the researcher should design quasi-experiments that can infer causal relationships even when the subjects being studied cannot be randomized. Finally, the original theory should be revised to incorporate inferences drawn from the new empirical information gained from the research.[66] The remainder of this article will more fully develop each one of these steps.

A. Developing Generalized Theories

The first step in measuring unobservable legal concepts is to develop a generalized theory that allows multiple inferences to be made about the unobservable construct.[67] All theories are not created equal; that is, a hierarchy exists among theories in which some are more elemental and apply to specific situations, while other more generalized theories, involve concepts that can be tested under multiple conditions. For instance, a researcher studying a new educational model could devise a basic elemental theory that a certain student will perform better on tests if the new model were in place because she has been receptive to similar educational models. Unfortunately, this theory provides little information because the results are very specific and is affected by the student's particular characteristics. Hence, inferences that can be made about the quality of the program in a larger population are limited. A more generalized theory addressing how the curriculum will affect the entire class becomes more interesting; it allows the researcher to asses the program using multiple operations, rather than a single test. Hence, if a convergence of data occurs (namely, that a majority of the class improves after being subjected to the treatment), a stronger inference can be made about the quality of the program. But the class may still have its own individual traits that can influence the results as well, and an even more generalized theory regarding the effect of the new educational model on the entire school can provide still more information. Similarly, generalized theories about law can provide the best information about the influences laws can have on a population by allowing more inferences to be drawn about the legal construct.

Legal academics are skilled at developing and debating legal theories, but often the theories tend to be elemental, in that they are limited to the application of specific laws or concepts. Professor Thomas Ulen has pointed out, however, that a common attribute in most sciences is that theories are universally applicable without regard to specific governmental

---

[65] Smith, *supra* note 19, at 396.
[66] These four steps are based closely on a five-step model developed by Greg Smith for construct validity research in psychology. *Id*. at 399
[67] Ulen, *supra* note 33, at 897-99.

institutions.[68] In addition to the multiple inferences that can be drawn from generalized theories, disciplines that rely on generalized theories also have the advantage of better communication and more focused study across boundaries; scholars are able to address common issues and share their findings without regard to their personal location.[69] In economics, for example, researchers around the world share the same basic theoretical principles---such as the rational-actor theory---that allow their hypotheses to apply generally.[70] Because economic research is often based on generalized theories, economists in India can share their research with Americans because both groups speak the same theoretical language.

In contrast, law does not have accepted general theories that transcend political boundaries and form a theoretical paradigm within which legal researchers can more easily communicate across borders. Rather, legal scholars focus on more elemental issues that affect their particular government.[71] More productive theories, however, address generalized concepts that are amenable to empirical measurement because they allow researchers to use multiple techniques to explore the theory. An empirical legal scholar studying an elemental theory about a single law passed in a small community can run only a limited number of experiments to explore the effects of the law. For example, the community can be analyzed both before and after the law was passed or the community can be compared with similar communities that do not have the same law (experimental design will be discussed more thoroughly below). In contrast, generalized theories about a class of laws allow multiple approaches for researchers: multiple communities with different laws within the class can be explored or the effect of that class of laws can be studied within the same community. While these same experiments could be conducted without a generalized theory, they would progress in a state similar to what Kuhn described as an immature science in that they would lack cohesion that could provide focus for future study. The theory could provide that cohesion and suggest potential follow-on research.

---

[68] *Id*. at 894.
[69] *Id*. at 894-99.
[70] *Id*. at 895. Under the rational-choice model for human behavior, social scientists assume that people accurately determine the risks and benefits of their actions and act accordingly. *Id*. at 886. This is the current paradigm within the social sciences and is the foundation for much of its current literature. *Id*. at 886-87.
[71] *Id*. at 897.

1. Drawing international interest in interesting topics

A general theory can draw more interest than an elemental theory, and can advance the current knowledge for the field more quickly and robustly.[72] By developing theories that address interesting and widespread issues, they are more likely to draw commentary, which can foster a developing community for scholars to knowledgeably debate each others' work. This community of critical discussion is especially important in law, which lacks significant peer-reviewed journals. In other fields, studies are vetted for publication by experts based on the quality of the research and the logic of the conclusions. And despite calls for a change, law currently lacks this filter. Hence, the only way under the current regime to ensure quality work is through discussion and debate. If a study produces controversial results, others can evaluate the methods and comment on the conclusions. This way, well-executed studies can be celebrated and cited, while lower grade experiments will be undermined by criticism and eventually relegated to obscurity. The best way to encourage the growth of these essential debates is for legal scholars to frame their theories in ways that are not limited to a specific statute or legal doctrine, but to question how the concept behind a law can result in observable behavioral alterations throughout a population.

In contrast to science, the elemental theories often investigated in law rarely attract interest across national boundaries.[73] While other social sciences benefit from diverse commentary from scholars internationally, legal debates tend to remain imprisoned within political boundaries. Professor Ulen argues that "there is no persuasive case for 'legal exceptionalism'---i.e., for the view that law is inherently different from other academic disciplines that characterize themselves as scientific."[74] The obvious benefit from garnering increased attention beyond national boarders is that more people can provide useful commentary, insight, and critical analysis of an idea. Legal experts in other countries may be able to provide differing perspectives on an issue and help provide a more nuanced view of the problem.

Psychology Professor Greg Smith describes how the psychological theory of self-enhancement (the tendency to concentrate on one's own strengths rather than weaknesses) was reworked based on information acquired through multicultural study.[75] Initially, psychologists believed self-enhancement was universal, but cross-cultural testing showed that

---

[72] *See* Epstein & King, *supra* note 1, at 896-97.
[73] Ulen, *supra* note 33, at 895.
[74] *Id*. at 899.
[75] Smith, *supra* note 19, at 398.

inhabitants of some Asian cultures failed to demonstrate the trait.[76] After further investigation, the researchers learned that members of these cultures did, in fact, self-enhance, but they exhibited the trait in a different manner: while members of individualistic cultures focused on individual behaviors, members of collectivist cultures placed more weight on collectivist behaviors (e.g., defending decisions made by the group).[77] The critical study of diverse cultures provided information that showed errors in the assumptions that supported the theory of the universality of self-enhancement in ways that domestic studies could not. The result of observing these varying conditions was a more detailed and complete understanding of the field.

This example also illuminates the nature of auxiliary theories: general theories depend on the validity of auxiliary, or supporting, theories,[78] and negative empirical data could be the consequence of faulty auxiliary theories rather than failures in the theory being examined. Theories that can be tested in other cultures can be used to undue this ambiguity by allowing international commentators to highlight the possible assumptions on which the theory is based. But despite the benefits of cross-cultural study, current legal research is limited almost exclusively to national issues and cases decided in United States courts. And while that research is certainly beneficial, like the situation with self-enhancement in psychology, some commonly held beliefs in law may be misguided because they are based on particular American qualities.

For example, most criminal law scholars agree that increasing mandatory prison sentences leads to decreasing unlawful behavior. As a result, when legislators became concerned about the spread of crack cocaine, they increased the mandatory sentences for selling crack as compared to powder cocaine. But it may be possible that Americans have a particular aversion to longer prison terms. Although the possibility seems unlikely, in other communities the very idea of being sentenced to spend time in prison alone may be such a significant deterrent that a mandatory minimum ten year prison sentence will do little more than a one year sentence. By performing cross-cultural studies examining the effect of longer prison terms on the sale of crack cocaine, researchers can determine whether the influence increased punishments have on unlawful behavior is a universal constant or if it is a particular cultural trait. This knowledge can then inform decisions about future sentencing changes in the United States as well as other countries.

---

[76] *Id*.
[77] *Id*.
[78] I. Lakatos, Lectures on scientific method, in I. Latakos & P. Feyerabend (Eds.) *For and Against Method* 19-112) University of Chicago Press (1999).

2. An example in constitutional law

The derivation and benefit of a generalized theory can be seen through an example posed by Judge Richard Posner, who suggested questions about constitutional law that he would like to see explored empirically.[79] The first question he posited was "What difference has it made for press freedom and police practices in the United States compared to England that we have a judicially enforceable Bill of Rights and England does not (or at least did not, before it became subject to the European Convention on Human Rights and Fundamental Freedoms)?"[80] Any researcher attempting to empirically answer this question will quickly confront the difficulty in measuring "freedom of the press." As a first step in empirically answering this question, then, one could pose a generalized theory that countries with explicit constitutional guarantees tend to express freedoms more vibrantly than countries that have not made express constitutional rights. Put this way, the question about the freedom of the press in the United States versus England becomes an elemental component of a more general question about constitutional law and its influence on societies.

Posing a theory in such a broad manner is helpful for several reasons. The first is the benefit gained from inviting a larger audience to comment. In addition to attracting First Amendment experts to the issue, scholars in various aspects of constitutional law may also become engaged in the debate. And as the study of this question draws more commentary, more ideas will be generated that help drive the field; critics may run tests in an attempt to undermine the theory, while supporters may try to replicate the outcome of a positive test or otherwise defend the theory. In the end, a rigorous debate can develop with perspectives beyond the confines of First Amendment debate and the theory will become more refined and more accurate as empirical knowledge is acquired.

Another significant benefit from addressing the elemental question about the First Amendment as part of a more generalized issue is attention from abroad. Because the theory does not specifically address any one nation, legal scholars in other countries may have input as well.[81] International legal experts may have the ability and experience to bring nuanced insights to the theory in the same way that cross-cultural experiments allowed psychologists to refine their theories about self-enhancement. For instance, researchers may compare not only the United States and England, but also China, which specifically restricts speech and

---

[79] Richard Posner, *The Problematics of Moral and Legal Theory* 156-57 (1999).
[80] *Id*. at 156.
[81] Ulen, *supra* note 33, at 895.

the press. Analysis can also be done on certain Middle Eastern countries that allow only state run press organizations. Perhaps, by empirically studying the varying relationships between governments and the local press (from constitutional freedoms to constitutional controls), scholars will be able to refine their understanding of the effects of constitutional rights in general.

Finally, the generalized theory about constitutional rights can create a unifying concept that can draw together studies that would otherwise run without direction. Studies of the Fourth Amendment's protection from search and seizure can be compared to the First Amendment's protection for the press. Also, studies conducted in other countries that explore their particular constitutional system do not have to stand in isolation, but can be contrasted with their American counterparts. And, like Kuhn's paradigms for science, the generalized theory of constitutional rights will suggest future empirical legal work to progress the current understanding of constitutional rights.

3. An example in intellectual property law

Intellectual property (IP) provides an example of a law that is derived from generalized theories about incentives and rights, and has been extensively debated without regard to national borders. The basis of intellectual property laws are generally attributed to one of two basic theories: natural rights or utilitarian.[82] The theory of natural rights, developed during the Enlightenment, states that IP rights are an inherent part of the laws of nature---creators should own their creations. In contrast, according to the utilitarian theory, IP is a legal right needed to provide an incentive for citizens to create; the public will forgo some of its ability to use new innovations by allowing its government to provide a limited monopoly right to creators, thereby providing a financial incentive for the creation. Utilitarians believe that without IP law, innovations will not be utilized at their most efficient levels. The American system of IP derived primarily from the utilitarian model, but debate continues as to the extent that the limited monopoly is necessary to drive invention.

Many of the most recent modifications to copyright law in particular tend to provide increasing protections for creators. For example, the Digital Millennium Copyright Act (DMCA) makes criminal tampering with effective encryption devises that protect copyrighted material, even if the encryption also prevents access to material that is not copyrighted or if the person wishing to gain access may intend a legitimate fair use of the

[82] Richard A. Epstein, *Liberty Versus Property? Cracks in the Foundations of Copyright Law*, 42 San Diego L. Rev. 1 (2005).

copyrighted work. The Supreme Court has also recently allowed a twenty-year extension to the length of most copyrights. Opponents of these laws have argued that these additional protections were unnecessary because the existing laws provided sufficient incentives for artists and other innovators.

It is impossible to directly measure the "incentive" created by IP rights, but because of the generalized theories on which they are based it is possible to infer the degree the laws influence creation by empirically exploring the multiple traits that the incentives influence. For example, because patent law has a standard duration for all inventions regardless of the development cycle of the product, it is possible to infer the influence of the law by comparing the frequency with which inventors patent in different industries. In the pharmaceutical industry, for instance, the life cycle of a product often outlasts the life of the patent thereby decreasing the financial benefits that the innovator can obtain by exploiting their monopoly rights. But in the computer industry, new innovations become obsolete long before the monopoly rights expire, so the inventor can optimize their profit for the life of the product. It is possible to compare the patenting practices in these two industries, in which IP law provides varying protection, to learn more about the actual incentives the law provides.

Cross-cultural studies of intellectual property could provide similar benefits. American IP laws are based largely on the belief that people are more willing to create if they receive a financial incentive for their work. But again, the fact that Americans seem to produce more when they are given stronger intellectual property rights may be a reflection of the country's reliance on capitalism to reward positive behaviors. In socialist societies, citizens may feel stronger motivations to innovate to provide greater benefits for the community, rather than for individual gain. By studying the influences of intellectual property laws in other countries, researchers can gain a fuller understanding of these effects.

B. Infer multiple critical hypotheses

Because generalized theories do not apply to single law, they allow researchers to infer multiple critical hypotheses to make more detailed inferences about an unobserved legal construct. Latent variables cannot be determined from testing one hypothesis alone; they can be measured only after detecting trends of convergence and divergence of multiple operations conducted to test multiple hypotheses.[83] As researchers accumulate evidence in support of a theory by testing the multiple hypotheses,

[83] Likewise, in similar situations econometricians assemble a set of several hypotheses that question multiple aspects of the subject.

confidence in the theory can increase.[84]

As a concrete example, consider how astronomers study an unobservable singularity, commonly known as a black hole. Singularities do not reflect light, radio, or other signals, so it is impossible to directly see or hear them. But astronomers have still been able to examine several throughout the galaxy. To do this, they first surmised the possibility of such celestial bodies based on theoretical considerations. They then could deduce that if such a body existed with a strong gravitational pull, then observable objects in the surrounding area should be influenced. Astronomers were then able to scour the skies to look, not for the holes themselves, but for their effects on surrounding bodies. When astronomers noticed that a star moved in an irregular pattern, they could hypothesize that the irregularity may have been caused by the gravitational force of an unobserved body---possibly a singularity. But even if the influence is the singularity, it is still impossible to determine exactly where it is, how large it is, or how strong the gravitational pull. If multiple observable bodies were affected in the same area, however, then it may be possible to "triangulate" details about the unobserved object without ever being able to detect it directly. For instance, if orbits for several different objects were pulled in the same direction but at different severities, it may be possible to infer an approximate location for the singularity. But, just as a theory can never be proven completely true, astronomers cannot truly confirm the existence of a singularity because it can never be directly observed.

Unobserved legal constructs can similarly be inferred based on their influence on observable operations. And while a single operation may indicate the possibility that the unobservable trait exists, only after testing multiple hypotheses is it possible to reliably triangulate the details. Moreover, by testing a larger number of hypotheses, more information about the unobservable trait can be gained. For example, a generalized theory may state that capital punishment deters violent crime, and a researcher may infer the hypothesis that a state that employs the death penalty should have a lower rate of violent crime then one without it. But just like a single irregular orbit does not provide enough information to study a singularity, the mere decrease in violent crime does not provide a full picture of the deterrent effect of certain punishments. The crime rate may have been influenced by a third factor, such as an up-tick in the economy that created more jobs for the unemployed who may have otherwise resorted to violence. Hence, to establish the validity of the deterrence theory, the researcher must also test hypotheses that economic or other factors influenced the crime rate. And although the theory can never

[84] Smith, *supra* at 19, at 397.

be proven completely true, as more hypotheses are tested, the validity of the theory can become more certain.

But this example also demonstrates the necessity for critical hypotheses. Because the deterrent value of capital punishment can never be completely established, a powerful set of hypotheses does more than simply reinforce a theory. A hypothesis that, if proven, supports an existing theory can provide more confidence in that theory. Conversely, a hypothesis designed to address criticism of the theory or help choose between competing theories can provide more valuable information; if such a critical hypothesis is supported based on statistical analyses, then the community learns that the theory may be false. But if the hypothesis fails, critics have one less avenue through which to attack the theory. Powerful hypotheses eliminate as many criticisms as possible, and a hypothesis that completely undermines a critique can provide the strongest possible evidence in support of the theory. Hence, a powerful set of hypotheses for legal concepts will suggest experiments that can help rule out as many criticisms as possible. For example, with regards to the theory about deterrence, a researcher may test the hypothesis that economic conditions in a state with capital punishment are actually responsible for any difference in observed crime rate.

1. Operationalization

Before hypotheses can be tested, however, the legal constructs being examined must be defined. Many terms have different meaning for different people; for instance, when in 1964 Justice Potter Stewart was called on to define obscenity under United States law, he famously wrote that he'll know it when he sees it.[85] But for an empirical legal scholar trying to study the effects of a new law or Supreme Court case on the amount of obscenity, a more useful definition is necessary. A more functional definition for research may be that obscenity contains frontal nudity or explicit sex acts. Based on this definition, anyone can determine if a film is obscene. While others may not necessarily agree with this definition, they will be able to understand and assess research that uses it.[86]

Clearly defining a concept in a way that it can be measured, a

---

[85] *Jacobellis v. Ohio*, 378 U.S. 184, 197 (1964) (Stewart J., concurring) ("under the First and Fourteenth Amendments criminal laws in this area are constitutionally limited to hard-core pornography. I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it, and the motion picture involved in this case is not that.")

[86] Bruce L. Berg, Qualitative Research Methods for the Social Sciences 25 (1998).

process called operationalization, is a vital component to quantitative science.[87] Operational definitions are descriptions of variables or constructs in terms of the specific validation tests used to measure them, rather than in terms of an intrinsic essence.[88] For example, weight may be defined as the result of putting an object on a scale and temperature may be defined as the reading on a thermometer.[89] In psychology, mental retardation is operationally defined as a score of 70 or lower on an I.Q. test.[90] Happiness can be defined in terms of facial expressions (such as smiling), tone of voice, and other observable characteristics; thus, if psychologists wish to measure whether certain treatment makes someone happy, they can count the number of times a person smiles in a given time-period after the treatment is administered.

Once a construct has been empirically studied, it may still be necessary to adjust the operational definition if the results do not correlate with expected operations. For example, the term "genius" may be defined by a certain IQ score, but experts also expect geniuses to score well on achievement tests in school. If studies were to find, however, that students with genius IQs performed poorly on achievement tests, then the term genius may have to be redefined using a measure other than IQ score. In law, if obscenity was defined as the proportion of a film's screen time devoted to showing nudity, but films that meet this definition of obscene do not also contain a higher than average number of sex acts, then the term obscenity may need to be redefined.

Although law schools do not currently provide much training on how to analyze complicated statistical problems, lawyers are particularly trained in how to define difficult concepts. When interpreting a statute, lawyers argue the meaning of particular words or concepts. In applying judicial decisions, lawyers must determine how to apply the idea put forth in previous cases to the facts of a new problem. The skills required to determine the precise definition of an ambiguous statute or the application of a complicated legal rule are similar to those required to define an abstract idea.

2. Hypotheses for the constitutional rights example

Turning back to Judge Posner's question about the degree of liberty provided to the American press due to the First Amendment and our theory

---

[87] *Id*.
[88] *Id*.
[89] *Id*.
[90] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, Text Revision (DSM-IV-TR)* 41-43 (4th ed. 2000).

that freedoms explicitly proscribed in a constitution will provide more liberty, one notes immediately the difficulty in directly measuring the term "freedom". Although many scholars have debated the true meaning of the word freedom, empirical research requires an operational definition. Hence, it is necessary to operationalize the term before hypotheses can be inferred to test the theory. One possible definition for freedom of the press is the proportion of stories and editorials that are critical of the present government versus the overall number of stories published. Conversely, in a country with less freedom of the press, reporters will be more reluctant to say anything negative about the presiding national rulers for fear of possible repercussions.

Once freedom is defined by the number of unfavorable stories towards the government, hypotheses can be inferred to test the theory. For example, one might hypothesize that the United States will exhibit more freedom than England because of the constitutional protections for the press. Hence, to test the hypothesis an experiment might analyze the proportion of critical stories in the American press versus the British media. But just as a single irregular orbit is not sufficient to identify a singularity, this single hypothesis cannot support the theory alone. To more completely test the theory, a researcher must develop several hypotheses that can eliminate potential criticisms and help choose between alternate theories.[91] For instance, a critic may point out that the proportion of critical stories may actually be a reflection of an unpopular regime or a general cultural attitude towards authority. It is thus necessary to devise several hypotheses that can help establish the theory.

A more complete set of hypotheses to test freedom of the press would include the one dealing with stories critical of the government, but may also include other hypotheses. One hypothesis may be that any difference in disapproving reporting is the result of an unpopular government. To test this hypothesis, a researcher may compare opinion polls or collect data from various time periods. If this hypothesis proves false, then this criticism has been undermined and it is more likely that the press freedom is a result of constitutional protections. Another hypothesis could be that any differences are actually caused by a more widespread social acceptance of criticism of the government; some societies may view critical analysis a positive trait for the press and will reward contrarian news outlets with increased readership regardless of the legal framework. To test

[91] This idea is similar to that developed in econometrics by Haavelmo in *The Probability Approach in Econometrics* (1944). He proposed that econometricians formulate a set of *a priori* admissible hypotheses to address a given theory. The set of hypotheses should be based on the existing on the theory at the time to help ensure that a proper hypothesis was included in the set.

this hypothesis, an experiment could be devised to survey different societies' view of the proper role of the press and compare these results with the presence or absence of an explicit constitutional freedom or even the existence of laws limiting freedom of the press.

The benefit of the more generalized theory about constitutional freedoms is that it will also allow comparisons between countries other than the United States and England. Another way to test the theory then is to compare countries with constitutional rights similar to the First Amendment to countries not just with no explicit right, such as England, but also to countries with specific controls over the media, such as Iran. Governments around the world exert varying degrees of control over their local media, and studies of press freedom in these different countries will provide a clearer and more complete picture of the influence of the First Amendment.

3. Hypotheses for the intellectual property example

A similar process can be used to infer hypotheses about the degree of incentives that intellectual property laws have over creation. First an operational definition must be devised for the term "incentive". In this case, incentive can be the number of patents and copyrights that are issued either by governments' patent or copyright offices.[92] Because some countries do not have IP or equivalent offices, another definition for international study may be the number of creative products, such as books, CDs, or technological developments that are produced for the market. While critics may disagree with these definitions, they will still be able to understand and asses the research.

After incentive has been defined, it is possible to infer a hypothesis to test the concept. A first hypothesis could be that a longer duration of the limited monopoly protection provided by copyright law will provide more incentive to create; if IP laws provide incentives to create, then stronger IP laws should provide increased protection and thus more incentives. One way to test this hypothesis is by comparing the numbers of copyrights sought before and after copyright protections were extended by twenty years in 1998 to include the life of the author plus 70 years. But this test alone is not sufficient to support or undermine the theory. Assuming the

[92] While trademarks are a form of intellectual property, and could potentially provide an incentive to create, they should not help define incentives because the justification for trademarks is not the same as that for other forms of intellectual property. Copyrights and patents are designed to form limited monopoly rights for the IP rights holders. In contrast, trademark law is designed to protect consumers from confusion between products. Even though less consumer confusion may lead to a larger incentive to produce, it more attenuated that the direct benefits derived from copyrights and patents.

number of copyrights issued increased after the passage of the extension, other factors may have influenced the change; for example, the DMCA, which criminalizes decryption devises used to pirate copyrighted material, may have increased the value of the copyright independent of the change in duration, or the two laws may somehow interact to produce results that neither would have alone. But even if the results show no change in the number of copyrights issued, those results are not definitive on their own. It is possible that the incentive provided by copyright law maximizes at 50 years after the author's death so the extension had no overall effect even though other changes in duration would have changed the results.

Hence, it is necessary to infer other critical hypotheses to test the theory. For example, production of creative material can be compared between the United States and China, which does not have the same level of IP protection. The generalized theory allows study in countries operating under various IP regimes to provide a more precise understanding of the incentives provided by the laws. Other beneficial hypotheses may address areas in which people create without IP protection in an environment in which protections may be available. For example, while academic journals sell their publications, the individual authors rarely receive a direct financial benefit from producing work. Similarly, a recent study explored aspects of the fashion industry in which piracy of designs is common, yet designers rarely seem to enforce their IP rights.

C. Testing the hypothesis package

Once a set of critical hypotheses have been inferred from the generalized theory, the hypotheses must be rigorously tested using appropriate research designs. Application of proper experimental design to empirical legal scholarship has been debated extensively. Although the details of research design are beyond the scope of this paper, for the sake of completeness this section will briefly outline some of the basic concepts.

Empirical legal studies are generally conducted for three reasons: to collect data, to summarize data, or to make descriptive or causal inferences.[93] Because the legal community performs many activities that produce huge amounts of raw statistical data, the mere collection of the data is useful; but researchers must compile and organize the information so it can be used for future analysis.[94] Summarizing the data in a comprehensive but manageable format is also beneficial so the information can be understood quickly. But the most beneficial purpose of empirical research is to draw causal inferences---that is, past outcomes can be used to infer

[93] Epstein & King, *supra* note 1, at 17.
[94] *Id*. at 23-24.

future results.[95] Scholars can use information about the past to make two types of inferences about future events: descriptive inferences (predictions made about an entire population based on data gathered from a smaller sample set) and causal inferences (determinations of the factors that influence a given outcome). A researcher may make a descriptive inference about future events after studying how the opening of a large factory affects a small nearby town. The results can be used to surmise how other communities will deal with nuisance disputes in the future.[96] A causal inference could be useful to determine whether affirmative action laws have been a contributing factor to any change in the number of black lawyers.[97]

1. Quasi-experiments

After the goal of a particular research project has been identified, an appropriate design can be selected. A research design is the initial plans that provide the structure for the project and are generally categorized into three groups: true (or randomized) experimental designs, quasi-experimental designs, and non-experimental designs. Research using randomized experimental design (the gold-standard in research design) is used to establish whether a given treatment causes a specific effect: i.e., if X treatment is applied, then Y should result. But to reinforce the analysis of the causal effect, the corollary must also be true: i.e., if X treatment is *not* applied, then Y should *not* result. To make these parallel determinations, subjects from a common population are randomly assigned to either a treatment or a control group; subjects in the first group receive the treatment, whereas subject in the other group do not. Because subjects are assigned to the two groups at random, they are assumed to be essentially the same, or equivalent. A randomized experiment provides evidence in favor of a hypothesis when the predicted result occurs more often in the treatment group than the control. These types of experiments have high *internal validity*; that is, the treatment employed was probably the cause of the effect observed.

But because it is often inappropriate or impossible to randomly assign subjects to study legal concepts, empirical legal scholarship generally employs types of *quasi-experiments*, devised by Professors Donald T. Campbell and Julian C. Stanley.[98] These designs resemble true

---

[95] *Id*. 29.

[96] *See* Gideon Parchmovsky & Peter Siegelman, *Selling Mayberry: Communities and Individuals in Law and Economics*, 92 Cal. L. Rev. 75 (2004).

[97] Epstein & King, *supra* note 1, at 34.

[98] Donald T. Campbell & Julian C. Stanley, Experimental and Quasi-Experimental Designs for Research (1963).

experiments, but without randomized assignment. The lack of randomization means that quasi-experiments generally have lower internal validity than randomized experiments; it cannot be definitively determined whether an observed outcome was caused by the study's treatment or by a threat to the validity caused by an unidentified influence present in one group but not the other.[99] Hence, in all quasi-experiments it is important to identify possible threats to internal validity.[100]

Researcher can control for threats to internal validity by utilizing one of three general strategies: relabeling, substitution, or elaboration.[101] Relabeling is used when the threat results from a mislabeled effect; that is, when the cause of an effect is attributed to the treatment when, in fact, it was influenced by a third factor.[102] One way to cure this problem, therefore, is to just relabel the effect to better describe the true influences.[103] Substitution, the second strategy for ruling out a threat, is when a measurement that is subject to the threat is replaced by one that is not; for example, by using a randomized experiment rather than a non-randomized design to overcome selection bias.[104] Finally, elaboration removes a threat to validity by adding additional comparisons to the experiment to try to disentangle the effect being observed from the threat.[105]

Several different forms of quasi-experiments have been devised with differing threats to the internal validity. Some of the most popular (and most valid forms) used in empirical legal scholarship involve "nonequivalent groups," in which intact groups, such as states, cities,

---

[99] The threats to validity arise from several forms of selection bias: history (some difference in the groups' histories changes how they react to the treatment); maturation (the groups' differ in their rate of maturation in relation to the treatment for reasons unrelated to the test); testing (somehow the pretest changed how the groups' approached the posttest); instrumentation (other influences on one group versus the other between the pretest and posttest, such as different observers); regression (different rates of regression to the mean between the groups, such as when one group has more extreme pretest scores, so they have further to regress); selection (the subjects have difference from one group to the other); mortality (different dropout rates between the groups); and interactions between these biases and selection. *See* Thomas D. Cook & Donald T. Campbell, Quasi-Experimentation: Design & Analysis Issues for Field Settings 51, 52 (1979); Campbell & Stanley, *supra* note 98, at 5.

[100] Shadish, *supra* note 4, at 13, 16.

[101] *Id.* at 89, 90.

[102] Charles S. Reichardt, *A Typology of Strategies for Ruling Out Threats to Validity, in* Research Design: Donald Campbell's Legacy 89, 91 (Leonard Bickman ed. 2000).

[103] *Id*. at 90-91.

[104] *Id*. at 92-93.

[105] *Id*. at 94. Reichardt also describes five methods of elaboration: demonstrate that the threat has no actual effect; subtract the size of the effect from the results; vary the size of the treatment effect; vary the size of the threat effect; and vary the size of both the treatment and threat effects. *Id*. at 95-108.

judicial jurisdictions, etc., are selected and presumed to be similar (but not equivalent as they would be if randomly assigned).[106] The Interrupted Time-Series with Comparison Series ("Comparison Series") approach compares similar groups both before and after a treatment, such as a new law or a new Supreme Court decision.[107] If the two groups are substantially similar before the law is passed, and the observations being measured are the same pretreatment but diverge posttreatment, it is reasonable to conclude that the law affected the measurement. Even so, plausible alternative explanations, such as a change in other local conditions or changes in other laws, should always be explored.

Professors Albert Yoon's and Tom Baker's empirical analysis of the effects of a New Jersey offer-of-judgment court rule is a recent example of a Comparison Series research design.[108] Under the New Jersey litigation rule that they studied, either party to a civil suit could offer a settlement to the opposing party; if the party receiving the offer refuses but goes on to lose the case, that party must pay all litigation expenses including attorney fees that were incurred after the offer was made.[109] Originally the rule set a cap on the attorney fees at $750, but the cap was later removed in amendments to the statute. The professors collected data from before and after the rule was revised to see if the increased cost-shifting associated with the amendment had an effect on settlement rates. To fully analyze the effects of the rule, the study analyzed data from in-court trials as well as from out-of-court settlements.[110] But because settlement information is usually not publicly available, Yoon and Baker arranged with a major American insurance company to use their confidential settlement data.[111] The study was designed so that the time period observed extended an equal duration before and after that change in law.[112] The dataset also included information from five other states other than New Jersey to serve as the control group---these other states did not change their offer-of-judgment rules during this time period.[113] Suits filed in New Jersey comprised nearly

---

[106] Social scientists use several different quasi-experimental designs: time-series, equivalent time-sample, equivalent material, nonequivalent control group, counterbalanced, separate-sample pretest-posttest, separate-sample pretest-posttest control group, multiple time-series, recurrent institutional cycle, and regression-discontinuity analysis. *See* Campbell & Stanley, *supra* note 98, at 5 (describing various quasi-experimental designs).

[107] Campbell, *supra* note 3, at 201.

[108] Albert Yoon & Tom Baker, *Offer-Of-Judgment Rules And Civil Litigation: An Empirical Study Of Automobile Insurance Litigation In The East*, 59 Vand. L. Rev. 155 (2006).

[109] New Jersey Court Rule 4:58.

[110] Yoon & Baker, *supra* note 108, at 159.

[111] *Id*. at 165.

[112] *Id*.

[113] *Id*. at 166.

20% of the data that the insurance company supplied.[114] The study concluded that removing the cap did not increase the number of settlements, but did decrease the average time to resolve a suit.

Yoon and Baker approached the data using what they called an economic approach.[115] If the researchers were able to run a randomized experiment, they would have randomly assigned subjects to control and treatment groups that were equivalent in every respect except for the variable being tested, in this case the cap on damages. But because it was impossible to randomly assign citizens to either a city with or without the new rule, the researchers compared suits from a control group that consisted of states with similar demographics and size to suits from New Jersey (the treatment group).[116] The control states did not have a change in their offer-of-judgment rules, but the treatment state did. In this way, the researchers made the quasi-experiment as close to a true experiment as possible by limiting the number of unexpected or exogenous factors that could influence the results.

But Comparison Series design is only one type of quasi-experiment. Other variations may be more appropriate depending on the situational limitations, such as if only one group can be observed,[117] or if only posttests can be conducted.[118] Similarly, other designs can be used with various analytical strengths and weaknesses, and using them in combination can strengthen confidence in causal inferences.[119]

2. Adopting research design to empirical legal scholarship

Professors Epstein and King have proposed a method for empirical legal scholarship based on the existing methodological rules applied in the social sciences.[120] They propose the following procedure for legal researchers: (1) develop research questions that contribute to existing knowledge and improve the real world; (2) formulate well-reasoned hypotheses about how societies will react if the theories are true; (3) bolster the theory by developing rival hypotheses that test the accuracy of the original theory; (4) use valid and reliable measurements used to test the hypotheses; and (5) select the appropriate observations to include in the study.[121] They also suggest that the legal academy should adopt an

---

[114] *Id*. at 169.
[115] *Id*. at 172.
[116] *Id*. at 173-174.
[117] This design is called a One-Group Post-Test design.
[118] This design is called a Posttest-Only Design with Nonequivalent Groups.
[119] Cook & Campbell, *supra* note 99, at 103-46.
[120] Epstein & King, *supra* note 1, at 11.
[121] *Id*. at 54.

infrastructure that is more conducive to empirical work by offering more courses to teach law students how to properly conduct empirical research, train law professors about empirical techniques and provide more resources to use these skills, and create more expert review for empirical articles submitted to law reviews.[122]

Professors Epstein and King also note the importance of articulating the methods used as precisely as possible. An experiment does not provide any value to the field if readers cannot decipher how a test was performed; results alone are meaningless without an explanation of how they were achieved. A test is useful when it rules out competing theories, but if it is not clear how a test is performed, hypotheses are not ruled out because readers cannot determine if the conclusions are based on the effects being studied or an unarticulated internal design flaw. Negative tests results may be caused by an incorrect auxiliary theory rather than a failure of the theory being examined. But if the methods are not clearly articulated, it is impossible to tell if the theory is wrong or if it is relying on a poor supporting hypothesis. Essentially the author of a study that does not fully explain the method used is asking readers to simply trust the author without ever verifying the conclusions.

3. Replicability

A related reason to fully articulate the experimental methods is so the test can be repeated by others to determine whether the study measured a real event or if another factor may have influenced the outcome. The results from a single experiment may reflect the phenomenon being studied but they also could reflect a chance outcome or a variable that the designer did not anticipate. For example, study results that are based on survey questions may be influenced by the medium; survey results may differ depending on whether they were gathered online, over the telephone, or on paper. More subtle details can also influence the data such as the color of the background or the text (participants in the study may have had a harder time reading light colored text against a light background). But it cannot be determined from the single experiment alone if the results are a reflection of the variables being manipulated, random happenstance, or a third unaccounted for factor. Hence, replication is essential. As a study is replicated several times, researchers can have more confidence that the outcomes are the result of the variables being studied and not just a random occurrence. But researchers wishing to repeat the study require detailed instructions to ensure their results are not influenced by other outside

[122] *Id*. at 114.

factors. More general descriptions of methods are more prone to lead subsequent researchers astray.

Professor Robert Thompson's empirical analysis of courts' willingness to pierce the corporate veil in corporate liability cases is an excellent example of a legal study with a detailed description of its methods.[123] Because of the interest in the topic and general confidence in the results, this study has become well-cited and has been referenced by both trial and appellate courts in federal and state cases. Thompson spends several pages of his article providing details about how he performed his research. The article first described the Westlaw searches he conducted to collect the relevant court decisions. He then explains that the original search results needed to be vetted to ensure that the cases that resulted from the search were actually related to corporate law.[124] Finally, the article elucidates the specific information gleaned from each case.[125] Overall, this description gives readers a detailed roadmap of the data collection process and how the results were analyzed.

But to make this study more easily replicated, it would be helpful to add still more detail about the methods, in particular, how the cases were vetted to ensure they focused on the correct topic. While the article precisely describes the specific searches conducted, it does not completely explain how the results were filtered. It states that cases "that did not address corporate law" were eliminated from the study, but before subsequent researchers can replicate this study, they would still need to know how Thompson defined "corporate law," and how it was determined what the case "addressed." The article does not explain whether his filter allowed only corporate liability actions or if it also counted insider-trading, antitrust, or litigation of other corporate laws. The methods section states that Thompson's research assistants made the final decisions about which cases to include in the study, but does not explain how they were instructed to make these decisions. If an appendix had been included with the exact filtering instructions, future researchers would be able to more closely replicate the study.

Unfortunately, even with this level of detail, follow-up studies may not arrive at the same results if the instructions allowed the assistants to exercise too much independent discretion to decide which cases to include; different readers may make different decisions about close cases.[126] And without more information about the screening process, readers cannot

---

[123] Robert B. Thompson, *Piercing the Corporate Veil: An Empirical Study*, 76 Cornell L. Rev. 1036 (1991).
[124] *Id.* at 1044-47.
[125] *Id*. at 1044.
[126] *Id*. at 1044 n.48.

determine if the results may have been influenced by the way the cases were gathered. Although this level of methodological detail may seem trivial, it can have profound effects on whether replication of the study will achieve the same results and ensure the reliability of the study's conclusions. As is clear in the next example, one solution for this problem is to allow subsequent researchers access to the final datasets.

Yoon's and Baker's study (described above), also lacks an important element for replicability. While using the information from a large insurance company seemed ideal because it is one of the few entities involved in many different lawsuits that also keep detailed information about the entire transaction, before allowing access to the data, the insurance company required that the researchers keep the company's identity secret.[127] So even though the methods were set out in detail (the article describes the exact information the insurance company provided, included a table that summarized the data that the insurance company provided, and described the specific statistical analyses performed)[128] because the dataset is confidential, subsequent investigators cannot attempt to replicate the study.

The inability to replicate has several implications. First, the study may have basic math errors that can not be checked. Clearly Yoon and Baker thoroughly tested and retested their data, but if subsequent investigators wish to rely on this study for future experiments, they will not have the confidence gained from analyzing the data themselves or the ability to use different statistical analyses on the data. For example, critics may wish to analyze some underlying assumptions on which the study is based. Supporters of the research may wish to build on this research by performing even more sophisticated statistical analysis than the ones run in this study. But currently law has not adopted an accepted norm to cope with confidential information. Confidential data is a common problem in medical and psychological experiments because medical data is often collected from individuals, and these fields have developed way to address these problems. Scientists have also developed methods of untying data from the individual from whom it was collected. Although peer-reviewed journals in most fields will not publish an article if the data is not available to other researchers, the data must be anonymous. This may involve assigning identification codes for each subject that only the original researcher can decipher.

Legal scholars performing empirical work can develop similar requirements. The subjects of a study, including corporations, should not be harmed by the experiment. The legal academy should therefore try to

---

[127] Yoon & Baker, *supra* note 108, at 165.
[128] *Id*. at 174.

develop a system, possibly based on the systems used in other sciences, which can safeguard subjects of studies while still allowing future study of datasets.

D. Theory Modification

Finally, once critical hypotheses of a generalized theory have been tested using appropriate research designs, the results should be used to modify the theory. Construct validation is an iterative process, in which theories are modified based on the latest empirical research and then future research can be conducted to test the new theory. Presently, however, empirical legal scholarship does not have a standardized system for incorporating new empirical data into existing theories. The result is that evidence contrary to a popular theory may be ignored or the academy may just be slow to recognize the new data. By incorporating new empirical data into generalized theories in a standardized manner, legal theories can evolve to become better predictors of behavior based on governmental policies.

Some philosophers of science believe the ultimate goal of scientific research is to find the definitive truth about nature.[129] Scientific research adds to our knowledge of the universe and, as science progresses, scientists come closer to the final truth. In legal studies, the goal may not be an ultimate true law, but a more modest attempt to find the best possible set of laws to govern a given society at a given time. Finding the best set of laws may then be accomplished by finding "true theories" about how law influences society. These true theories can direct lawmakers about which laws to pass to correctly apply a policy.

But, as noted above, empirical research should not be considered the final step in determining new policies. The quantitative analysis should be used only to inform new policy decisions. The generalized theories can be used to provide guidance when lawmakers approach new issues, but should not be used as the final solutions for problems.

1. Incorporating empirical results into the theory of constitutional rights

We have already considered a theory that explicit constitutional rights result in more freedom than if the rights were not included in a government's foundational document. We then inferred the hypothesis that if the theory is true, then the press should exhibit more freedom when in a country where it receives constitutional protections. After running several

[129] *See, e.g.,* Popper, *supra* note 7, at 50; Larvor, *supra* note 10, at 102; Smith, *supra* note 19, at 400.

quasi-experiments testing this and rival theories, we are now faced with new empirical data that either supports, undermines, or alters the original theory. We can now consider how to modify the theory based on the new information.

If the data confirms the hypothesis and has ruled out several alternate theories, then the theory has been supported. Although a theory can never be proven undisputedly true, the experimental results add confidence that it is accurate. This theory can now be used to inform future policy considerations. For example, if a nation was considering drafting a new constitution or new constitutional amendments, the drafters should be advised that if they include certain protections for citizens, the population is likely to act more "free" than if those protections are left to be set out in statutory provisions.

But even in these circumstances, researchers should not rest on their laurels. More research can still be conducted to further refine the theory or rule out other criticisms that had not yet been addressed. In addition, the hypotheses discussed thus far have only addressed freedom of the press. Future research can address other constitutional rights, such as a right to privacy, a right to counsel in criminal proceedings, a right to due process, etc.

If the experimental results undermine the validity of the theory, however, then the theory should be changed to reflect the negative data. If the results show that the press does not exhibit more freedom despite the presence of constitutional protections, then researchers should consider several alternative explanations. First, the operational definition of freedom may not have been correct. Counting stories that criticize the government may not have been an accurate way to measure freedom. Alternatively, it could be that the constitutional theory is correct, but the experiments relied on incorrect auxiliary theories. For instance, it may be that the experiments relied on theories about statistical techniques that are not accurate in this context. In any event, it is necessary to more fully explore the results and change the general theory to reflect the new data.

Most likely, however, is that the data will provide inconclusive results. The data may show that constitutional rights provide more freedom, but only with regard to certain rights or in certain cultures. If a culture is intrinsically adverse to criticism of authority, protections for the press may not produce any change in behavior. Similarly, a right for free speech may provide more robust results than a right not to quarter militia in one's home. By exploring these possibilities, the theory will become more refined and provide a better understanding of the field.

2. Incorporating empirical results into the theory of intellectual property rights

A similar procedure can be used to assess the generalized theory about the incentives provided by IP law. If it turns out that stronger IP laws result in an increase in the number of copyrights and patents, then those results reinforce the theory that IP creates an incentive to create. If, on the other hand, the experiments provide contrary results, then it is necessary to reassess the theory. It may be that incentives should not be measured by the number of copyrights and patents issued, but rather by the number for which people apply. Or, perhaps, it may be better to use several measurable indicators to get a fuller picture. Negative results may also mean that the theory being studied relied on faulty auxiliary theories, such as financial gains are the only motivations that IP law provides or that creators desire. Finally, it may be that the theory itself is incorrect and that IP protections do not have an influence on innovations.

But again, it is most likely that the results show that the theory is partially correct, but needs refinement. For instance, it may turn out that the motivation provided by IP law experiences diminishing returns. That is, a certain level of protection will spur creation, but the marginal benefits decrease after a certain level of protection is provided.

CONCLUSION

By incorporating concepts of measurement developed in social sciences, empirical legal scholars can better assess the impact of the unobservable constructs that underlie modern legal theory. Specifically, the notion of construct validation can be adopted from psychology to law to allow a standardized approach to the modification of legal theory to reflect updated understandings of how laws impact societies based on empirical observations. These modified theories, which are based on quantitative observations, can then better inform policymakers' qualitative decisions.