Evaluating Judges

Gregory Mitchell

Much of the interest in empirical studies of judges lies in the comparison of actual to ideal behavior. When we ask what makes a good judge or whether a judge rightly decided a case, we implicitly compare the judge’s decisions to a normative standard. In some instances, the content of the normative standard is uncontroversial and its application straightforward. Hence, a trial judge who sentences Black criminal defendants to longer terms of imprisonment than White criminal defendants, all other things being equal, departs from the normative ideal that application of criminal laws should be color-blind (e.g., Blair et al., 2004; Pruitt & Wilson, 1983). Or the circuit judge who votes to uphold a state law banning abortions on grounds that Roe v. Wade was wrongly decided departs from the normative ideal that inferior court judges should follow the Supreme Court’s constitutional decisions (for evidence on how circuit judges view their obligations within the legal hierarchy, see Klein, 2002).

In many instances, however, the content of the normative standard chosen for comparison purposes proves controversial or hard to delimit and its application unclear. Is the judge who applies a rule of evidence according to its plain language a “better” judge than one who applies the rule to achieve its purpose when following the text would contravene the purpose? Is the judge who deviates from the law to correct a disparate impact the law is having on minorities, say, in sentencing, a bad judge? Is the circuit judge who narrowly construes the Supreme Court’s abortion cases a better judge than the circuit judge who broadly construes this precedent? Does it matter if the former judge favors this minimalist approach due to pragmatic commitments rather than political preferences? Is there even a neutral approach to construal of Supreme Court precedent that can be labeled the way of the ideal judge?
Any empirical study that attempts to address these or other questions bearing on the competence of judges confronts three basic tasks: (1) specifying a defensible normative standard as the baseline for comparison; (2) converting the standard into testable form and judicial behavior into measurable units; (3) interpreting the results of any comparison to draw appropriate conclusions about the descriptive-normative gap. I consider complications at each stage in the comparison process, with illustrations from existing studies of judicial competence and behavioral decision theory studies from psychology, which examine the gap between actual judgment and decision-making behavior and norms of rational judgment and decision-making. I contend that we presently know very little about the degree to which judges depart from normative ideals because political scientists focus too much on the Supreme Court and normative standards unique to that court, psychologists focus too much on juries, and legal scholars focus too much on the normative side of the descriptive-normative comparison, but also because few studies clearly specify and justify the norms they seek to test and then carefully work through the complications of such testing.

I also consider whether we avoid these empirical and normative difficulties by embracing character-based or cognitive styles approaches to good judging. While those approaches do not avoid all of the usual problems and introduce new problems of their own, empirical studies derived from dispositional theories of good judging are likely to lead to new insights about judicial behavior and to a greater appreciation of the importance of legal structure relative to personal characteristics in determining normative compliance.

Knowledge about the descriptive-normative gap in judicial behavior is important for what it may reveal information about the lawfulness of particular cases but more generally for what it

---

1 I use the term “competence” to refer to a judge’s (degree of) compliance with a normative standard (see Stein, 1996, for a similar usage). The term is meant to have the same meaning here as the terms “rational” or “irrational” do within psychological studies into the concordance of actual judgment and choice with norms of rationality.
may reveal about the legitimacy of the legal system and ultimately the rule of law. Describing judicial behavior in relation to norms of proper judging holds an important place within debates about the proper role and power of the federal judiciary relative to other branches of the federal government and relative to state governments. As Sherry (2005) argues, the rise of the view of federal judges as just another political actor within the government surely contributes to elite, and perhaps popular (Schep & Lyons, 2001), discontent with judicial review, or at least provides cover for those elites seeking to change the balance of power (see Sisk & Heise, 2005). Characterizing judges as unelected, unaccountable political actors rather than experts in the administration of law takes the sting out arguments to limit judicial review and to give greater power to the legislature and executive. Likewise, to the extent that the legitimacy of the courts and their ability to motivate voluntary compliance with the law derives from the popular view of the courts as detached from ordinary political processes (Tyler & Mitchell, 1994), evidence that judges are just political actors in robes should undercut the legitimacy of the courts. However, these arguments should lose steam if existing studies into the descriptive-normative gap are far from compelling, as I contend they are.

From a prescriptive perspective, reliable knowledge about the descriptive-normative gap is needed to understand where judicial incompetence exists so that we may attempt to correct it. If judicial behavior systematically deviates from legal norms in some domains, then steps may need to be taken to rein in this predictably unlawful behavior. If judicial behavior commonly, but randomly, deviates from legal norms in other domains, then steps may need to be taken to

\[ \text{2 Within psychology, prescriptive models are distinguished from normative and descriptive models of judgment and decision-making. Normative models supply the standards or rules for evaluating judgments; descriptive models compare judgments to the normative models; prescriptive models specify the conditions under which judgments can be brought more into line with the normative models (see Baron, 2004).} \]
eliminate chance influences in these domains. If judicial behavior proves unresponsive to prescriptive interventions in these domains, then the legal norm should be reconsidered on grounds that ought implies can.

These are important questions that empirical researchers should be addressing. Nonetheless, it may appear that my goal here is to sow doubt about the ability of empirical research to provide satisfactory or useful answers about judicial incompetence. That impression would be a by-product of my trying to point out many of the difficult questions that researchers must address and how they will always have to choose a research strategy than can be second-guessed. Studying the descriptive-normative gap in judicial behavior is a very hard thing to do well, and it inevitably requires compromises to make a study manageable and its results interpretable. Many researchers no doubt struggle with these difficult choices, but I fear that too often expediency takes precedence over relevance. Addressing the limits of the normative implications of one’s study up front, and recognizing that even the best study of the judiciary can address only narrow questions about judicial competence, would seem to be obvious suggestions, but perusal of the political science literature on judicial behavior suggests it is less obvious than

3 These random deviations constitute mere performance errors, whereas systematic deviations would suggest an underlying incompetence (Stein, 1996). Both may need attention, but likely different sorts of attention. 4 Anyone who has thought through how to judge judicial worth, as all conference participants surely have, has probably already felt the sense of futility I sometimes feel when thinking through these issues. Indeed, I may be missing some of the biggest anxiety-producing questions. (For an extended discussion of the difficulty operationalizing good judgment for assessing real world judgments, see chapter 1 in Tetlock, 2005.) My meager and real goal is to provide a platform for discussion at the conference of what issues we should be most concerned about in these evaluative studies and whether there is any agreement on how best to go about studying judicial competence.

Perhaps, as Dave Klein has suggested, the answer is to focus on uncontroversial common denominators of good judging:

If we were to ask members of the legal community to describe the kind of reasoning that an ideal judge would employ to decide cases, the details would differ from one respondent to another. It seems likely, however, that all would agree on some fundamental desiderata: for instance, that the judge take into account all and only truly relevant evidence, weigh the evidence appropriately, be aware of the considerations that are playing a role in his or her decision, and not allow emotion to rule reason (Klein, 2005, p. 1716). I am skeptical that widespread agreement of even this kind is possible (at least with respect to the last two desiderata), but I am prepared to be persuaded otherwise.
one would think. At the same time, legal scholars and judges (e.g., Edwards, 1998) who dismiss the empirical studies of judges as hopelessly flawed and evidencing serious misunderstandings of what judges do risk being in the position of dismissing systematic empirical research for their own unreliable intuitions and observations and putting the competence of judicial behavior outside the domain of empirical testing.

By suggesting the many different standards that might be used to examine judicial competence and by suggesting the limits of the usual comparison between “attitudinal-type” models and simple “legal models” of judicial behavior, I hope to encourage the debate to move beyond characterizations of judges as driven primarily by their ideology or by the law. At all levels of the judiciary, judges’ political attitudes undoubtedly play a role in some of their judgments and decisions, but many other aspects of a judge’s psychological make-up and of the social-psychological setting in which the judge acts will be influential as well, with some of these influences pushing the judge toward idealized normative behavior and some pushing the judge away from judicial ideals. Gaining a more sophisticated understanding of when and why judges deviate from the various judicial norms should aid in the important task of developing prescriptive models of judicial behavior.

I. The Limits of Our Present Knowledge on the Descriptive-Normative Gap

One need not endorse a skeptical view of the limits of social scientific knowledge to endorse the view that we presently know little about the size of the descriptive-normative gap in judicial behavior. Political scientists, who by far have conducted the greatest number of

---

5 I suspect, from my conversations with political science and law colleagues, that my characterizing the debate this way will exasperate some because not all political scientists embrace the strict attitudinalist view and very few legal scholars embrace a strict legalistic view. Segal and Spaeth’s influential framing of the debate this way, however, surely has framed the larger debate, for better or worse, and their strong skeptical view of the power of the law to constrain (or the slightly modified strategic view in which other political institutions but not the law constrain judges) now regularly serves as the default position within empirical studies of judicial behavior.
empirical studies of judicial behavior, focus a disproportionate amount of their efforts on appellate courts, and the U.S. Supreme Court in particular (e.g., Friedman, 2006; Klein, 2002). Even if the strongest claims of the “attitudinalists” and “quasi-attitudinalists” (Baum, 2006, p. 7) are true about the role of personal policy preferences and strategic preference-maximization behavior in shaping the decisions of Supreme Court justices, there is little reason to believe that the strong version of either version of the attitudinal model generalizes to judges on inferior courts. Although lower courts are studied much less commonly, some empirical evidence indicates that legal norms influence judges on lower courts to a greater extent than on the Supreme Court (Klein, 2002; Rowland & Carp, 1996), and social psychologists would argue that the situation of the judge exerts significant influence on judicial behavior. From this perspective, studies of supreme courts have little external validity: one cannot assume that judicial actors in very different situations (especially who do not control their docket in the way the Supreme Court does) will act the same; equivalence must be empirically demonstrated.

Psychologists, for their part, simply have not conducted much empirical study of judicial behavior. Whereas mock jury studies make up a high percentage of all law and psychology empirical studies, studies of judicial behavior are infrequent. Consider, for instance, that Ellsworth and Mauro’s chapter on “Psychology and Law” in the most recent edition of the Handbook of Social Psychology contains no discussion devoted specifically to studies of judicial decision-making (Ellsworth & Mauro, 1998). Thus, while psychology may have much to offer the study of judging (Wrightsman, 1999), psychologists have had little to say about it.

Legal scholars spend enormous amounts of time telling the courts how they should decide cases—indeed, normativity is almost a prerequisite to publishing in a law review (Rubin, 1997)—but relatively little time examining what, beyond the reasons given in opinions,
motivates courts and whether courts can in fact follow the norms that legal scholars recommend to them. When legal scholars do consider the descriptive side of the descriptive-normative comparison, they tend to favor qualitative studies that identify descriptive-normative gaps in particular cases (e.g., Barnett, 2005) or lines of cases (e.g., Barnett, 2004). While qualitative studies may generate important hypotheses for future testing and rich stories about how judges do or do not comply with legal norms, such studies can provide only limited evidence about the systematicity of this behavior (i.e., their conclusions are of dubious generalizability\(^6\)) and cannot test competing causal hypotheses due to sample restrictions (Büthe, 2002; King et al., 1994).

A variation on the last point is worth emphasizing: beware of vivid anecdotal evidence of alleged judicial misbehavior (of the kind found, for instance, in Boot, 1998). Segal and Spaeth (2002) begin the revised edition of *The Supreme Court and the Attitudinal Model* with discussion of the Supreme Court’s decision *Bush v. Gore*, which they take to be a recent example of judicial policy making. Baum (1997), on the other hand, begins his review of judicial behavior research with anecdotes about conservative Supreme Courts deciding *Roe v. Wade* and then affirming *Roe* in *Planned Parenthood v. Casey* to suggest that judicial attitudes cannot do the explanatory work often assigned to them.\(^7\) Neither book bases its main conclusions on anecdotal evidence; rather, each understandably uses the anecdotes for narrative purposes. But the contrast illustrates the key point: empirical debates cannot be decided with anecdotal evidence, as each side will be able to find a seemingly endless supply of supporting stories. That is not to say that exemplary judges or exemplary cases may not be important sources of information about the characteristics and behavior of good (or bad) judges and the institutional

---

\(^6\) A deviation from the norm may be the result of measurement error or random error, such as scoring error, an error due to a momentary lapse of attention, or just rotten luck (Clemen, 1999) (citation omitted). Qualitative studies may not be capable of distinguishing random errors from systematic deviations from norms due sample restrictions.

\(^7\) Segal and Spaeth (2002, p. 289-292) assimilate *Casey* to their attitudinal model and thus do not treat it as dissonant evidence.
features that may promote or inhibit good judging; of course, reaching consensus on who are exemplary judges returns us to the original problems of identifying the norms to use for evaluating judges and measuring the exemplary qualities.

II. Alternative Normative Perspectives on Judging

Normative standards derive from analytical frameworks that enjoy authoritative status, either by general acceptance because the framework seems to work well or because it has been justified or promulgated by a source to whom deference is given (Baron, 2004). Some frameworks impose order, such as the rules of arithmetic; other frameworks are supposed to help us achieve desired goals efficiently, reliably, or accurately, such as expected utility theory for the maximization of utility or scientific principles for the development of reliable empirical knowledge. Some normative systems, including systems of social and legal norms, are not easily reducible to a single clear, consensual goal or function (Elster, 1989), and some normative systems remain quite controversial, as with the debate over the propriety of norms of rationality for evaluating human judgment and choice behavior (e.g., Cohen, 1981). For the most part, empirical researchers take norms as identified by others (usually philosophers, statisticians, game theorists, or legal scholars) and apply them to actual judgment and decision-making behavior without wading into the foundational normative disputes.

The potential norms for evaluating judicial behavior may be divided into five categories: social, moral, legal, coherence, and efficacy norms. Social norms, in this context, involve inter-judge comparisons: Did Judge X behave as the median judge behaves in such a case? Are the judges of the Ninth Circuit outliers in their treatment of Establishment Clause cases? Although

---

8 Others may parse the norms differently; this particular categorization, which is used for organizational purposes, holds no special importance in the paper.
some judicial studies do discuss social norms within courts, few treat deviation from judicial social norms as a reason in and of itself for criticizing a judge’s behavior.9

Moral norms, as I use the term here, encompass any norm derived from a theory of justice or ethical obligations. Legal scholars often use moral norms to evaluate individual judicial decisions (see Posner, 1999), but these studies typically dispute the normative status of the law by recourse to moral arguments rather than describe judicial behavior. Several empirical studies examine the degree to which parties or citizens in general perceive judges and the courts to be acting justly or to be allocating outcomes justly (e.g., Tyler & Huo, 2002; Tyler & Mitchell, 1994). Procedural justice norms are an important source for evaluation, but they typically are used in connection with efficacy norms as discussed below. Other than procedural and distributive justice norms, few empirical studies of judging evaluate judges against moral norms as defined here (except with respect to selected decisions), and little more will be said directly about moral norms or social norms.

Legal norms are norms established by an authoritative governmental body, including constitutional and statutory law and judge-made common law, to bind the institutions and persons governed by this body. The power of courts to clarify, establish or alter legal norms complicates matters from an empirical perspective, for it means that in some instances it is legitimate for a court to ignore, alter, or invalidate one norm to resolve conflict with another norm (as may occur through judicial review or common law development) or to embellish a norm (as may occur through legislative or constitutional interpretation). Uncertainty about what counts as law or binding legal norms likewise complicates matters (Hart, 1961), as we have seen with recent disputes about the role of foreign law in U.S. courts (see Rahdert, 2007).

9 The few works that do evaluate judges in relation to social norms usually involve norms of secrecy and civility (e.g., Gaffney, 1994). Baum (2006) argues for greater attention to collegiality as an important judicial motive, which might lead to a greater emphasis on social norms and the degree to which judges comply with them.
Accordingly, a secondary set of legal norms has developed to govern interpretive practices and to establish a hierarchy of precedent and deference to guide superior and inferior courts, future panels of the same court, and branches of the government as they sort out the content of legal norms and the obligatory status of these norms. For instance, the norm of rationality review within constitutional law provides that a legislative act with some conceivable purpose behind it should be upheld by the courts.\(^\text{10}\) It is often this secondary set of interpretive and role norms that serve as the primary legal norms in comparisons of actual to ideal judicial behavior (e.g., Segal and Spaeth, 2002, operationalize the “legal model” of judicial behavior as requiring fidelity to plain meaning, legislative or constitutional drafters’ intent, and precedent).

Coherence norms, including norms of rationality, require logical consistency and coherence in the formation and ordering of beliefs and preferences, whatever those beliefs or preferences may be:

> technical discussions of rationality generally adopt a *logical* conception, in which an individual’s beliefs and preferences are said to be rational if they obey a set of formal rules such as complementarity of probabilities, the sure-thing principle or independence of irrelevant alternatives. In the laissez-faire spirit of modern economics and decision theory, the content of beliefs and of preferences is not a criterion of rationality—only internal coherence matters (Kahneman, 1997, pp. 105-106).

\(^{10}\) See, e.g., FCC v. Beach Communications, Inc., 508 U.S. 307, 315 (1993) (“On rational-basis review, a classification in a statute such as the Cable Act comes to us bearing a strong presumption of validity, and those attacking the rationality of the legislative classification have the burden ‘to negative every conceivable basis which might support it.’ Moreover, because we never require a legislature to articulate its reasons for enacting a statute, it is entirely irrelevant for constitutional purposes whether the conceived reason for the challenged distinction actually motivated the legislature.” (citations omitted)). Whether rationality review actually operates in this minimal, deferential form is the subject of debate (see Farrell, 1999), but it does supply a normative baseline.
“The goal of this coherence metatheory of judgment is to describe and explain the process by which a person’s judgments achieve logical, or mathematical, or statistical **rationality**” (Hammond, 2000, p. 53).11

Rationality as coherence requires that judgments exhibit (1) resistance to logically irrelevant features of the situation (e.g., post hoc judgments about the reasonableness of a course of conduct should be immune to whether the course of conduct caused no harm or severe harm, given that this information was unavailable when the conduct was undertaken) and (2) appropriate sensitivity to logically relevant features of the situation change (e.g., base rate information should be properly assimilated into a Bayesian calculation) (Kahneman, 1991).

Furthermore, (3) judgments should lead to a coherent and consistent web of beliefs (i.e., a change in one belief may require that other related beliefs be altered) (Tversky & Kahneman, 1982).

Coherence norms evaluate behavior solely in terms of its coherence and consistency, with no necessary connection to real world success or empirical accuracy:

> It may come as a surprise to the reader that rationality does not directly imply accuracy and vice versa, but brief reflection shows that this is the case. Rationality always operates in a closed system; given the premises, certain conclusions follow if a rational reasoning process is followed. When the reasoning process satisfies a logical test, the system is termed coherent, and that is all it is and all it claims to be (Hammond, 2000, p. 53).

This does not mean that behavior within a coherence metatheory is purposeless nor that purposes or beliefs are irrelevant to rationality in this system. It means only that goals and beliefs are to

11 Technically, legal norms are a subcategory of coherence norms because they constitute a closed system of rules to be followed without regard to some independent measure of the quality of system outputs. The goals are determined by the body that promulgates the legal norms, and the legal norms are designed to produce outputs in line with the desired goals. Legal norms are thus analogous to scientific norms: scientific principles for conducting empirical research, if followed, should lead to reliable empirical knowledge; the scientific status of research is determined by its fidelity to scientific principles rather than the reliability of the knowledge produced (Kitcher, 1993). Likewise, fidelity to legal norms should lead to the kinds of outcomes lawmakers desire. I treat legal norms separately from other coherence norms because of their special importance for judicial evaluations.
be defined by the individual, and then behavior is examined to determine whether it is logically consistent and coherent in light of these goals and beliefs and whether beliefs are appropriately updated in response to new evidence.

In contrast, efficacy norms test judgments and decisions against empirical reality. The most common efficacy norm used in psychological research on judgment and decision-making is the correspondence norm: “The goal of a correspondence metatheory is to describe and explain the process by which a person’s judgments achieve empirical accuracy” (Hammond, 2000, p. 53). Thus, whereas psychologists operating under a coherence metatheory of judgment may examine whether causal attributions seem to follow norms for unbiased testing of causal hypotheses—with no attempt being made to assess the accuracy of these attributions—psychologists operating under a correspondence metatheory of judgment may examine the circumstances under which causal attributions tend to be more or less accurate (Funder, 1999). Correspondence norms, in theory, provide the simplest source of comparison standards because they provide an objective baseline for comparison and hold obvious attraction for evaluating the decisions of judges. Of course, in many instances, reality is not so easily measured (e.g., truth is a correspondence norm for judging the behavior of judges in criminal trials, but it is often very hard to determine truth in criminal trials).

Efficacy norms focus on the outputs of a system to test whether the system is functioning as it is supposed to according to a functional analysis or division of labor (whereas coherence norms focus on fidelity to the prescribed rules within the system). Thus, if our perceptual system leads to effective navigation in our environment, then the perceptual system receives high marks under efficacy norms despite evidence of systematic perceptual illusions (Funder, 1999; Hammond, 2000).
A variety of possible efficacy norms for judicial evaluation exist, most obviously norms of accuracy and error avoidance: how accurate is the judge’s fact-based rulings and empirical predictions and how does the judge resolve evidential ambiguities: does she err in the direction desired by the system (e.g., significantly more Type II than Type I errors in verdicts in criminal cases and slightly more Type II than Type I errors in verdicts in civil cases (where the null hypothesis is not guilty/not liable), or substantially more Type II errors on threshold rulings in civil cases, such as rulings on motions for failure to state a claim)? Or we may ask whether judges succeed at obtaining compliance with the law in their adjudications, however that goal is achieved. Or the related question of whether parties are satisfied with a judge’s rulings.

Note that the norms discussed above derive from analytical frameworks developed independent of what happens in actual cases. That is, the norms derive from some theory or observable consensus about what makes judgment and decision-making behavior good or bad or what goals judges should strive to achieve. Alternatively, one could reason backwards from cases where, say, litigants (or others) express low and high levels satisfaction to search for patterns of judicial behavior associated with those reactions and try to discern what norms, if any, the judges tend to follow or what cognitive and motivational qualities these judges tend to possess. Such an empirics-based approach to norms of good judgment has not, to my knowledge, ever been undertaken in a formal or systematic fashion with respect to judicial behavior, but historical and anthropological studies of courts may contain elements of this approach (consider that Alexander George (1980) and Irving Janis (1982), for instance, developed their important models of good political decision-making through historical studies of flawed and successful executive decision-making). Most evaluative studies of judicial behavior begin with a norm or theory of good (or bad) judgment and compare judges to that theory or
norm, but that need not be the only way to approach the descriptive-normative gap (I return to this topic briefly in the concluding section).

A. General Complications

The process of choosing a normative standard and applying it to judicial behavior presents predictable complications, some more difficult than others. First, the researcher must decide whether to take an internal or external normative perspective. From an internal perspective, the researcher asks the judge (or otherwise determines, perhaps through the judge’s opinions) what goal he or she was trying to achieve, selects the proper evaluative norm given that goal, and compares behavior to the norm to measure degree of success or compliance with the norm.12 An external perspective asks what goal the judge should be trying to achieve and applies norms proper to those goals; no effort is made to gain judicial input on proper normative standards. An external perspective may be justified from a meta-theoretical stance about the proper role of judges within the legal hierarchy or from the perspective of interested parties subject to the authority or influence of the courts (e.g., litigants may hope that judges’ decisions are legal, just, and accurate). Or the external perspective may be chosen simply to test some theory about the degree to which judicial behavior, whether intentionally or not, accords with particular normative standards.

The choice of norms is crucial, because different norms will often point to different conclusions about competence. A judge presented with inadequate evidence at trial may irrationally convict the truly guilty defendant, while another judge presented with this inadequate evidence may rationally acquit the truly guilty defendant. The first judge did a good job under the correspondence standard but a poor job under the coherence standard, and conversely with

---

12 We may question the sincerity of judges’ stated goals, and the possibility of insincerity does complicate the interpretation of results obtained from an internal perspective: if empirical tests reveal a large gap between actual and ideal behavior under the judge-chosen norm, then the gap may be due to insincerity or incompetence.
respect to the competence of the second judge. Hence, the choice of normative criteria controls the assessment of the judge’s competence on particular tasks.

It may be that the judge was capable of providing the normative response the researcher is testing for, but the judge had different values or concerns than the researcher assumed or intended and these other concerns overwhelmed the judge’s interest in apply the norm of the researcher’s choice. Here, it is not that the judge has misunderstood her task or role nor that biases drove the judge’s behavior, but rather that there is a mismatch between the task or role as the researcher perceives it and as the judge perceives it. In some instances, appearing logical, consistent, or fair-minded may be less important to the judge than achieving some other goal (e.g., moral and public policy concerns may override base rate information in determining liability), and, in other instances, the researcher may have mistaken beliefs about the judge’s goals or beliefs in light of the larger social setting. In these latter cases, where the researcher simply misconstrues the judge’s beliefs or goals but the judge’s behavior is logical in light of her true goals or beliefs, then we cannot say that the behavior is irrational under a coherence standard if that is our normative perspective of choice.

“Unifunctionalist tunnel vision blinds the research community to empirical and normative boundary conditions on basic effects. Inconvenient though it is, people are multifunctional entities that demand cumbersomely complex explanations (Tetlock, 2002, p. 469). Thus, a

---

13 Windschitl and Wells make this argument: “Consider . . . the general implication of a system of justice that could find people liable, based on the balance of probability that derives merely from membership in a group or class. Would it be good social or legal policy, for instance, to consider a man to have harassed female coworkers merely because he is a member of a corporate organization in which 80% of the males have harassed female coworkers. Our point is that it would be inappropriate to use such base rates for these types of judgments (Windschitl & Wells, 1996, p. 41).”

14 As an alternative way to understand the possible problem of confounding between seemingly biased behavior and underlying values and beliefs, consider how the interplay between cognitive bias and different aversions to the types of verdict errors may play out at trial. Imagine a criminal case with Judge 1 who considers conviction of an innocent person to be a mortal error that must be avoided at all costs and Judge 2 who has aversions of equal
narrow view of behavior may reveal departures from the coherence definition of rationality, but a
cleaner view of the same behavior may reveal a larger purpose that makes this lack of coherence
look instrumentally rational and that confounds conclusions that the irrationality was due to
cognitive bias.\textsuperscript{15}

Existing discussions of judicial competence typically employ externally determined
measures of competence, rather than internally-determined measures based on the judge’s self-
chosen goals or norms. Because we are usually interested in how judges fulfill their roles within
the larger legal-institutional setting, it is fair to impose values and goals on the judge that he or
she might not endorse or that might even be contrary to the judge’s own goals or values—
although sharp disparities between internal and external normative perspectives should be rare if
the judge deems her legal system legitimate and the researcher accurately perceives the goals of
the legal system.\textsuperscript{16} Nevertheless, a finding of external incompetence does not necessarily mean
that the judge was incompetent from an internal perspective; it only means that the judge failed
to meet an externally-imposed normative standard. Indeed, a judge may be judged externally

\textsuperscript{15} With surprising frequency within behavioral decision theory, researchers simply assume that study participants
would endorse a particular goal with a particular set of normative standards, without justifying this external
normative perspective. Which may lead to premature judgments of incompetence:

Even when people appear to be making systematically biased judgments or irrational decisions, it
is likely that they are trying to solve some problem or achieve some goal to the best of their
abilities. The behavioral researcher is well advised to look carefully at his or her research
participant’s behavior, beliefs, and goals to discern “the method in the apparent madness” (Hastie,

\textsuperscript{16} If the judge takes, that is, something like Hart’s (1961) “internal point of view” and feels obliged to obey and
uphold the law. That does not mean the judge will do so with complete success, given psychological limitations as
well as evidential and normative ambiguities, but the judge internalizes the legal system’s rules and tries to follow
them.
incompetent precisely because the judge is quite competent at achieving a goal that is internally desirable but undesirable from an institutional, external standpoint.

Second, because there may be dispute about the goals that should be served by certain tasks or because the judge may need to serve multiple goals, the researcher may need to employ multiple, possibly conflicting normative criteria or justify her choice of some normative criteria to the exclusion of others. Otherwise, the research may be dismissed as irrelevant or be disputed on grounds of incompleteness. Absent an agreed ranking of values and norms, however, contending factions may exalt or impugn judicial competence simply by invoking norms against which judges fare better or worse.

Third, we come to what may be the single most difficult problem in attempting to study compliance with legal norms (and which may explain the preference for indirect tests of norms that I discuss below): even if there is no dispute over the proper norms, there may be dispute over, or uncertainty about, how to operationalize them for testing purposes. If efficacy norms of accuracy and party satisfaction are employed, for example, then the researcher must come up with a way to measure accuracy and outcome satisfaction. Experiments provide a particularly attractive setting for addressing this problem, because they allow the researcher to vary the dimensions of hypothetical cases to examine whether fact-finder behavior conforms to different formulations of the norms and allows the specification of true and false results (as, for example, in tests of false confession detection).

In some cases, it may be impossible to formulate workable tests of goals, which may result in the selection of a second-best goal. Indeed, we might prefer that truth be the ultimate

---

17 Or it may be that different legal systems embrace different goals, and it may be impracticable to considerable all of the different goals within a single study if each goal requires a variation in design to test for competence.
18 For instance, Feigenson (2003) criticized the Sunstein et al. (2002) study of punitive damage awards (mock jurors served as the primary subjects, but one study in the collection examined judges) for its emphasis on optimal deterrence as the goal of punitive damages to the exclusion of a retribution as a goal.
touchstone in fact-finding, but we may have to settle for a coherence measure of competence over a correspondence measure because we lack a reliable measure of truth in many real cases. Of course, while many of the rules of evidence can be seen as rules designed to lead to accurate decision-making, other rules endorse values other than truth and prevent the introduction of evidence that might improve accuracy. We may thus favor a coherence standard that asks whether judges follow the rules of evidence because we believe that these rules are the most feasible and manageable means to truth approximation in light of epistemic constraints and because of the other values contrary to truth detection that the trial process must serve. Which brings us to a more detailed discussion of legal norms.

B. Legal Norms

Many empirical studies of judicial competence employ “perfect lawfulness” as the normative standard: Do extra-legal factors affect judicial decisions that ideally should turn solely on what the law and legally-relevant facts of the case dictate? The numerous studies by political scientists into the correlation of Supreme Court justices’ political attitudes with the direction of case votes on judicial decisions employ this normative standard (e.g., Unah & Hancock, 2006). Most prominently, Segal and Spaeth (2002) contrast their attitudinal model of judicial decision-making with what they call the “legal model,” which is supposed to represent how a law-bound ruling would be made (through fidelity to text, drafter’s intent, or existing precedent), and they find that justices’ different political attitudes relate significantly to different patterns of votes on cases.

There are considerable limitations to the use of perfect lawfulness as the normative standard. First, the norm is typically just “in the air” and does not serve as a source for direct comparison. In these studies, the normative position of perfect lawfulness at best serves as the
null hypothesis, and any measured extra-legal variable that is statistically related to case outcomes leads to a rejection of the null hypothesis. In Segal and Spaeth (2002)’s main tests of the attitudinal model of judicial decision-making, there is no direct comparison between judicial decisions and the predictions of any specific instantiation of the “legal model.” Rather, Segal and Spaeth, after finding that ideology can explain some significant percentage of variance in judicial decisions, simply assume that this finding contradicts what a legal model would predict.

Of course, to the extent a legal model would predict absolutely no influence of justices’ attitudes, then the legal model is contradicted, but I am aware of no legal model that would predict that liberals and conservatives would perfectly agree on the controlling legal norm in all Supreme Court cases. Indeed, I am aware of no single legal model or set of coherent legal norms that all conservative scholars and judges would agree on as the correct approach, nor that liberal elites would similarly agree on, and I suspect that many politically conservative and politically liberal elites would disagree over the proper outcome in many Supreme Court cases from even a narrowly self-interested, purely political perspective.19 In other words, support for the attitudinal model certainly casts into doubt “perfect lawfulness” as an accurate description of judicial behavior, but the degree of deviation from this vague standard goes unspecified, and, given competition over what perfect lawfulness requires across cases, interpretations of a particular outcome as reflecting inappropriate policy preferences or principled disagreements between liberals and conservatives as to what the law requires in particular cases.20

19 To the extent that ideology and normative perspectives are confounded (e.g., conservatism is likely to be positively correlated with acceptance of textualism as a legal process norm, whereas liberalism is likely to be positively associated with purposivism as a legal process norm), the correlation between outcomes and ideology does not compel the conclusion that legal norms fail to constrain decisions. Correlation cannot establish causation: the conservative may conveniently favor textualism to advance desired policy ends, or the conservative may endorse textualism as part of a sincere set of beliefs about the proper role of the judiciary; partial correlations can suggest the proper causal model but cannot establish that ideology causes a strategic choice of normative perspectives.

20 A direct comparison would compare actual votes to the votes predicted by a legal norm. Very few studies undertake this direct comparison, perhaps because “it is very difficult to operationalize a quantitative empirical test
This last point is not a minor quibble with political science studies of the Supreme Court. If a conservative judge can choose among legally-principled courses C1, C2, and C3 to reach the conservative result that he politically prefers, but if the judge feels constrained by legal norms to choose C1 over the other two options, then this constraint is significant if writing the opinion pursuant to C1, as opposed to pursuant to C2 or C3, has effects on the lower courts and later Supreme Court decisions (see Richards & Krtizer, 2002; see also Friedman, 2006).

Second, because these studies almost invariably test for an extra-legal influence rather than test for actual compliance with a legal norm, a finding that the variable of interest is not significantly related to judicial behavior tells us little about judicial competence other than that this variable, in this study, did not exert any influence. It may be the case that judges were acting pursuant to the normative standard or that the wrong extra-legal influences were investigated.21

Third, dependent variables (i.e., the judicial behavior of interest in extra-legal influence studies, such as votes to grant certiorari or votes on the merits of a case) are often operationalized in categorical terms that may be too insensitive to capture the influence of extra-legal factors or, if a significant influence is found, may obscure the real nature of this influence. For instance, judicial outcomes may reflect a wide range of political positions, from very liberal to very conservative, but Segal and Spaeth (2002, Ch. 8) categorize Supreme Court decisions as liberal or conservative, and the decisions they make regarding what constitutes a liberal or conservative decision reflect particularized versions of liberalism and conservatism (e.g., pro-federal-government rulings are generally deemed conservative, but of course some varieties of

---

21 As I discuss below, when testing legal proscription norms, which forbid consideration of some specific factor(s), a finding of no influence from these factors may vindicate the judge’s conduct relative to the legal norm. But tests for extra-legal influences will often not provide a direct test of compliance with other types of norms, as I discuss below.
conservatism would probably not uniformly endorse such rulings). That is not to say that
dichotomous dependent measures are not appropriate for testing some deviations from legal or
other norms; just that one needs to be sensitive to the loss of sensitivity that accompanies such
choices.

Fourth, many legal norms are process norms as opposed to substantive, or outcome,
norms (e.g., rules of procedure and evidence at the trial level, or interpretive norms for statutory
and constitutional construction and standards of review at the appellate level). Process norms
specify the range of permissible considerations and at best forbid some categories of outcomes
that could be reached only through reliance on certain sorts of impermissible evidence or
considerations, such as using lotteries or coin flips to assign responsibility for torts; process
norms narrow the range of permissible outcomes but do not compel particular outcomes. Unless
one takes the position that it is impossible to test compliance with process norms (a proposition
that psychological studies of the rules of evidence have falsified, see, e.g., Wistrich et al., 2005)
or that process norms have no important constraining force (an empirical claim itself), then the
overwhelming focus of political science studies on case outcomes as opposed to decision
processes greatly limits the utility of these studies of the descriptive-normative gap in judicial
behavior (Richards & Kritzer, 2002).22

A related problem is the requirement that a legal norm lead, a priori, to definitive
predictions for empirical testing purposes. This requirement is, of course, reasonable from an
empirical perspective, but its implementation has been problematic. As Segal and Spaeth (2002,
p. 59) correctly point out, a legal norm that cannot tell a judge how to act in a particular case

22 In some domains, process norms dominate legal scholars’ normative discussions. For instance, process norms
dominate discussions of constitutional interpretation (i.e., norms for how to make constitutional decisions rather than
norms about the specific decisions that should be made), making their relative neglect within empirical studies all
the more unfortunate given the great interest in these empirical studies in constitutional decisions.
cannot be the subject of empirical testing (nor can it have normative force) (see also Segal & Spaeth, 1994), but a legal norm need not compel a particular decision on the merits to have some constraining force (stated counterfactually, would the same evidence have been considered, would the same rulings on all issues have been made, or would the decision have been written in the same way had the legal norm not existed?). How an opinion is written may have more lasting influence (on the behavior of other courts as well that of lawyers and parties) than the particular outcome of a case; so it is no small matter if norms constrain opinions without altering outcomes (Richards & Kritzer, 2002). Whereas legal scholars may give too much explanatory weight to the reasons stated in opinions, the tendency to ignore the reasons stated and focus on case outcomes and external motives greatly restricts the usefulness and acceptance of empirical judicial studies by judges and lawyers.²³

An aside regarding the great emphasis within political science judicial studies on political attitudes as the key extra-legal influence on judges, a topic more properly addressed by other conference participants. Given the public policy issues often at stake in appellate cases, particularly constitutional cases, the emphasis within political science studies of judging on personal policy preferences as a likely motivating factor in how cases are decided is quite understandable. But political ideology or attitudinal orientation, as presently measured by political scientists (typically from the party of the President or the Senators in whose state the judicial vacancy is being filled; see Sisk & Heise, 2005) is unlikely to capture other important, enduring extra-legal influences. Basic assumptions about human behavior and the role of internal and external forces in determining behavior, about the ability of the market to correct

²³ Stated another way, if one of the goals of empirical judicial studies is to predict how judges will act and rule under specified conditions rather than assess normative compliance (see Leiter, 1997), then disregard of the text of statutes and cases constitutes a curious omission unless the prediction game is confined only to how a particular court—most notably the Supreme Court—will rule on a case once the judges hearing the case are known. Attorneys cannot play the prediction game in that truncated manner because they must often advise clients before the ideology of the whole series of judges who may ultimately hear a case is known. Accordingly, attorneys must look to existing case law and the text of statutes and regulations to predict how any generic judge would rule and advise their clients accordingly. Until studies demonstrate that looking to these documentary sources of the law is pointless to predict judicial behavior outside the Supreme Court, attorneys will continue to do so for lack of any better guidance on predicting how the law will be applied. Furthermore, when the prediction game is really a multi-level prediction game with tremendous uncertainty (trial judge A will rule this way, but appellate judges X, Y, & Z are likely to rule that way assuming X, Y, & Z are assigned to the case…), it is unclear whether any rational attorney, ex ante, should look more to judicial ideology than the law to advise her client.
inequalities and errors of judgment and choice, and about the harm of different types of errors may exert much more pervasive influence on judicial decisions than policy attitudes (see Tetlock & Mitchell, 2007), though no doubt there will be a correlation between conservatism-liberalism and various assumptions about human behavior and the power of markets. From a political psychology standpoint, we should focus not simply on political values but causal beliefs as well as determinants of public policy positions, not to mention a host of more mundane motives and influences (see Baum, 2006; Schauer, 2000). Of course, how to measure these other extra-legal factors creates a new large problem.

These concerns suggest that it may be wise to look for alternatives to treating perfect lawfulness as the null hypothesis and then declaring the legal model invalid when significant deviations from the null are found. In general, direct tests for compliance with legal norms, rather than indirect tests using the “perfect lawfulness” standard, will provide the information we are most interested in from an evaluative perspective. The nature of the direct test depends on the nature of the legal norm being tested. One way (and surely not the only way) to classify legal norms for purposes of testing is with respect to their injunctive content: (1) proscription norms: certain kinds of considerations must not be taken into account or certain processes or outcomes cannot be chosen under specified conditions; (2) permission norms: certain kinds of considerations may be taken into account or certain processes or outcomes may be chosen under specified conditions; (3) prescription norms: certain kinds of considerations must be taken into account or certain processes or outcomes must be chosen under specified conditions; (4) ordering norms: if multiple considerations or outcomes are prescribed or permissible, prioritize the considerations or outcomes in a particular way; (5) decision norms: decide a case in a particular way in light of permissive and prescriptive considerations.

Proscription norm tests will look very much like many of the extra-legal influence studies found within psychology that examine whether juries (and occasionally judges) can disregard inadmissible evidence as required by the rules of evidence (e.g., Wistrich et al., 2006).
Proscription norms can be found, of course, as the appellate level as well (e.g., rules about the record on appeal). The difference from the generic perfect lawfulness test is that the improper extra-legal influence is specifically defined by the specific proscription norm, and the norm typically specifies the conditions under which a particular consideration/outcome/process is not appropriate.

Permission norms vest judges with discretion to decide the propriety of certain outcomes or considerations with broad guidance as to the range of acceptable discretion, and, as such, these norms pose difficult testing problems for they require an empirical measure of acceptable discretion. One way to test for compliance with permission norms is to test whether discretion is exercised in a biased fashion. Thus, if judges with sentencing discretion impose higher sentences to Black than White defendants, all else being equal, then the judges exercise their discretion in a biased way. These tests require, however, another norm to resolve whether the bias is itself non-normative, as with a proscription norm against taking race into account in sentencing. If younger judges appointed by Democrats grant summary judgment more frequently than older judges appointed by Republicans in sexual harassment cases (e.g., Kulik et al., 2003), all else being equal, then it appears that one set of judges is using its discretion in a biased way, but it is not clear which direction constitutes the non-normative response. Another approach to permission norm tests would be to have experts agree on the appropriate range of permissive outcomes in a set of cases and test judges against these benchmarks.

Prescription norms allow point-specific tests of normative compliance. This category of tests includes testing for enforcement of bright-line rules, particular those triggered by objective conditions such as deadline rules. Because tests of prescription norms will typically yield clear
results (assuming the prescription norm can be operationalized and compliance with it measured without much controversy), detection of deviation from these norms should be straightforward.

Ordering norms will also allow for straightforward tests in many instances. Lower court obedience to Supreme Court precedent rather than contrary circuit precedent and preference orderings for types of evidence in certain types of disputes (e.g., statutory interpretation cases) will often yield clear results (of course, there may be dispute about whether the proper ordering norm is being applied). In other instances, such as implementation of the _Chevron_ norm of deference to administrative agency interpretations of statutes where the interpretation is reasonable, will prove more difficult because implementation of the ordering rule is based on subjective determinations about a case. In such instances, we must revert to looking for systematic disparities (tied to types of cases/values at stake, types of parties, or other situational or judge-specific factors) or deviations from expert orderings.

Decision norms specify particular outcomes when triggering conditions are met. The more objective the trigger and the conditions for determining whether the trigger is pulled, the easier the decision norm will be to test. Some decision norms function as constraints on the authority of the judge, such as minimalist or incrementalist norms about case law development or constitutional interpretation, and the focus will be on testing whether judges abide by such constraints. In tests of norms such as these, the key will be ranking possible decision outcomes in terms of restraint to monitor for compliance (or ranking for allegiance to text, intent, or purpose depending on the nature of the decision norm being tested).

From a psychological perspective, we should expect the greatest deviation to occur with permission norms (that assign discretion to judges and thus invite individual differences in
application), where triggering conditions for norms are most ambiguous,\textsuperscript{24} or where accountability is the least (e.g., supreme courts reviewing constitutional issues) or possibility for detection lowest.\textsuperscript{25} If the deviation appears to be in one direction across judges, then we should look for problems with how the norm is presented to judges (e.g., perhaps the text of a rule contains an unnoticed confusion) or for some systematic situational influence driving behavior in that direction. If the deviation appears bi-modal (or some other systematic but not unidirectional pattern), then it is likely that an individual difference variable or some co-varying situational influence is driving the deviations. If the deviation appears to be random, then it may be that the norm is difficult to apply in some subset of cases, this subset of cases is randomly distributed across judges, and, for this subset, some irrelevant situational or individual difference variable is having an influence that it does not outside this subset of cases. Random deviation may also represent the operation of a norm that is hard to implement for some reason (e.g., some complex rules of evidence may be difficult to implement consistently under the time demands of making real-time rulings on motions at trial).

\textbf{C. Coherence Norms}

There are, of course, many coherence norms other than legal norms that we may want judges to follow. Axioms of rational choice, rules of logic and probability, hypothesis-testing rules, and Bayes’ formula for updating beliefs are the most common coherence norms employed in behavioral decision theory studies.\textsuperscript{26} Many legal norms, such as interpretive rules and the

\textsuperscript{24} Arguably the central mediating variable in studies of extra-legal influences is the degree of discretion or subjectivity permitted or perceived with respect to any given judgment or decision, yet this variable typically goes unmeasured.

\textsuperscript{25} Here and in other places in the paper, I fail to provide citations from the psychology and political science literature to support the discussion because I did not want to further delay getting the paper to you. My apologies.

\textsuperscript{26} Although probability and hypothesis-testing rules ultimately serve the purpose of making accurate and reliable judgments, they are still types of coherence rather than correspondence norms, just like the rules of logic or rationality, because they do not guarantee correspondence to reality. If one complies with rules of probability and
rules of evidence, clearly allow room for other coherence norms to operate. In general, the more closely a judge follows norms of rationality, logic, and probability, the more consistent and predictable her behavior should be and the more likely the judge should be to maximize whatever values the judge seeks to maximize.

In the abstract, the notion that judges and jurors should act rationally when evaluating the law and evidence in a case is uncontroversial, and therefore proof from psychology of apparently rampant irrationality understandably commands the attention of legal scholars (as Saks & Kidd, 1980, noted many years ago). In some ways, this evidence should be more troubling to positive political theorists and others who advocate rational choice models of judicial decision-making as explanatory models of actual judicial behavior (e.g., Epstein & Knight, 1998; Schanzenbach & Tiller, 2006) than to scholars examining judicial behavior for evidence of a descriptive-normative gap. For there are a host of other norms that we may care equally about from an external perspective that do not require strict compliance with norms of rationality. Indeed, to the extent systematic deviations from rationality are caused by cognitive and motivational mechanisms that have evolved or are used to serve alternative goals than those served within the closed rationality system under observation (e.g., efficiency in processing at acceptable levels of accuracy (Hammond, 2000) or attributing greater causal responsibility to persons than rationally justified to encourage care and deter excuse-making (Tetlock, 2000)), then we should expect deviations from rationality to accompany satisfactory performance under alternative normative standards.

Nevertheless, whether judges, given their experience, training, and the decision aids available to them, comply with norms of causal hypothesis testing, probability and rationality is
of interest from an external normative perspective because this information may lead to prescriptive models of judging. We now know that several predictable judgmental biases can be debiased (Heath et al., 1998; Larrick, 2004), and we know that the structure of some environments ameliorate or exacerbate some biases (see Mitchell, 2002, 2003). We also know that there are individual differences in rationality, a topic addressed by Klein for the workshop. Identifying the norms that judges have the greatest difficulty following and identifying the characteristics of more and less rational judges may lead to feasible prescriptive models for legal reform.

Tests of norms of rationality and probability will look like many of the tests of legal norms, because all are coherence norms. Thus, if a norm of rationality directs the judge to ignore outcomes and focus only on ex ante information available to the parties to assess their behavior, we will test whether the judge can ignore these improper influences (see Rachlinski, 1998). Where the norm directs that certain information should be considered, such as base rate evidence in assessing the probative value of evidence, we will test whether the judge properly uses this information. Experimental tests allow the cleanest tests of judges’ ability to engage in analytical thought and comply with norms of rationality, but testing judges outside their natural environments may lead us to worry more than we should about the harms associated with laboratory normative deviations (Gigerenzer, 2006). Thus, experimental tests should be seen as providing an important first step toward understanding the rationality of judges and identifying possible normative deviations, but that step should be followed by studies assessing how institutions and individual judges adapt, if at all, to correct for such deviations.
D. Efficacy Norms

If one is interested in judges primarily as one of several interdependent actors within the governmental division of labor, then efficacy norms may be of particular interest for evaluative purposes. For instance, if a key role of judges within the larger institution of government is to induce voluntary compliance with the law to ease pressure on executive enforcement of the law, then testing judges for their ability to persuade parties to accept legal decisions and voluntarily obey the law become important efficacy norms for evaluative purposes (hence the popularity of measuring perceptions of procedural justice and its relation to perceptions of the legitimacy of the legal system and voluntary compliance with the law; Gibson, 2006; Tyler & Huo, 2002). Likewise, the function posited for constitutional adjudication by the Supreme Court will determine the efficacy norm one would choose to evaluate constitutional decisions.

The empirical and theoretical importance of efficacy norms is that they take the focus away from whether the law can constrain judges from turning their own preferences into law and refocuses attention on how effective judges are at achieving whatever goal is sought, by whatever means. Thus, a judge who possesses preternatural lie detection skills (see O’Sullivan & Ekman, 2004) and employed those skills to achieve highly accurate results in trials, while completely ignoring the rules of evidence, would be an excellent judge under a correspondence norm. The results one obtains from an efficacy norm study will only be as compelling as the argument one advances for the propriety of that norm, but, given agreement on the norms, efficacy norms arguably provide the most important information of any of the evaluative comparisons that can be conducted because such tests directly link performance to desired outcomes.
Once one satisfies oneself that the efficacy norm chosen is defensible and worth studying, then one confronts the task of operationalizing it and coming up with a criterion measure. Once outside experimental settings, where the experimenter controls the levels how the comparison to the norm will be conducted and measured (e.g., asking judges to rate known true and false confessions for truthfulness; cf. Kassin et al., 2005), a backwards or known outcomes approach for testing compliance with correspondence norms in trial settings may be wisest. In this approach one identifies cases where the values of the output variable of interest are known (e.g., innocence or guilt as determined with a high level of certainty based on DNA evidence or corroborated confessions by others (Garrett, 2007), or the outcomes of predictions made previously by judges), and then one examines the conditions under which judges made accurate or otherwise efficacious decisions. One must take into account, of course, that legal norms may constrain judges from giving certain types of evidence, at certain stages of a case, the weight the evidence might deserve under a strict correspondence test. Thus, such studies may reveal places where correspondence and coherence norms work at cross-purposes.

Efficacy norms have been relatively neglected compared to coherence norms, and legal coherence norms in particular. Yet getting things empirically right is what we often care most about in judging, and at times there will be objective (or at least inter-subjectively reliable) measures of efficacy. Appellate judges often make predictions about the empirical consequences of choosing one legal interpretation over another, and the most dire predictions are typically found in dissents, which makes their predictions testable in some instances (we should expect dissenters whose parades of horribles never materialize to claim that conditions changed, making predictions no longer applicable; Tetlock, 2005). For instance, Albiston and Nielsen (in press) tested the Supreme Court’s empirical claim that rejection of the catalyst theory of fee awards in
civil rights cases would not discourage the filing of suits by surveying public interest organizations, and their findings cast doubt on the accuracy of the Court’s claim. Or we can treat judicial decisions as natural experiments and examine the effects of these decisions on other courts, other institutions, and the pattern of lawsuit filings (e.g., examining the effects changes in the standards for admitting expert evidence; see Vickers, 2005).

III. Dispositional Approaches

I want to close by considering briefly an alternative to examining specific judicial behaviors for compliance with normative standards. A dispositional or character-based theory of good judging posits that the possession of certain traits or qualities is sufficient for judicial competence (e.g., Solum, 2003). Accordingly, judges should be evaluated for good character rather than good decisions.

This dispositional approach, without more, is little more than unproven tautology: wise judges will issue wise decisions (because the trait of wisdom means acting wisely). To move beyond tautology, some justification needs to be provided for giving the theory normative weight. That is, because a dispositional theory of judging does not directly invoke independently validated or justified normative standards (as coherence norms in particular have), the theory must somehow gain normative force. Normative authority could arise from agreement, as with a “thin theory of judicial virtue,” in which only uncontroversial traits that are desired in all judges are chosen (Solum, 2003). Or its normative status could arise from showing that the dispositional theory of judging derives from another theory that already enjoys normative status (e.g., Solum derives his virtue-centered theory of judging from virtue ethics). Or the theory could be justified through empirical observation (which I believe is the basis for Sherry’s (2005, 2006) views on the traits and experiences that appellate judges should have or should not have—
namely, in this last respect, experience as an academic): if empirical observation reveals that judges who possess qualities X, Y, and Z reliably produce excellent decisions, then these qualities should be favored. The problem with this last approach, of course, is that it simply moves the normative question to another point in the equation: we will need justified normative standards for deciding whether the decisions of these judges were indeed excellent.\(^{27}\)

One interesting feature of character-based approaches to evaluating judges is that, if one accepts the character traits chosen, then the evaluative focus shifts to whether judges possess these traits and away from performance. The task becomes measuring the traits of existing and potential judges to evaluate them for good and bad judicial character.

From an empirical perspective, the measurement problems associated with dispositional approaches to judicial evaluation are severe for judicial traits of the kind Solum advances (virtues and vices),\(^ {28}\) but much less so for judicial traits of the kind Sherry advances (at least for the experience-based factors she argues for as proxies for the more abstract character traits she ultimately favors). Consider a popular judicial trait within some dispositional approaches: good judgment, which may go by a variety of labels, such as wisdom or pragmatism (see Sherry, 2003).\(^ {29}\) There is no single theory of judicial pragmatism or judicial wisdom, and certainly no well-validated measure of pragmatism or wisdom. It may be difficult to distinguish a pragmatic from a textualist outcome in a particular case (because, as Posner, 2005, notes, at times the

---

\(^ {27}\) Or we might ask what kinds of judges seem to act in ways that are contrary to existing normative theories, using the character of these judges as negative evidence. For instance, the attitudinal model research would suggest that judges who hold less extreme ideological views should rule very differently than ideologues, and, if moderates are more constrained by the law (i.e., if their decisions are less predicted by ideology), then political moderation should be favored to the extent legal constraint is desired.

\(^ {28}\) Solum accepts that virtues of the kind he advocates are not quantifiable or objectively measurable (Solum, 2005a). Elsewhere he offers templates of the judicial qualities to look for (Solum, 1988, 2003), but in my reading of Solum’s writing in this area I have not found much specific guidance on how exactly to identify these qualities in particular judicial candidates.

\(^ {29}\) Posner (1995) provides one of the more explicit defenses of pragmatism, even if the content remains somewhat vague, though Posner seems to argue less that some are predisposed to be pragmatic judges and more that all judges should strive to be pragmatists.
pragmatic judge will adopt the textualist outcome), but it will be near impossible to distinguish a pragmatist from a textualist from biographical material about judges (though Solum, 2005b, disputes this with respect to his proposed virtues and vices) (and, in any event, “clinical” assessments of character are considerably less reliable than those based on validated assessment tools, none of which exist for pragmatism). If a judge’s opinions serve as the source for making determinations about the judge’s pragmatic character, then we return to our tautology problem.30

Another interesting feature of these theories, and the most problematic from a social-psychological perspective, is the assumption that judges with ideal “judicial character” are more likely to follow the law strictly, wisely interpret and apply the law, or, perhaps, do justice (whatever those terms imply) than those without this character. While it is certainly the case that individuals differ in important ways that may relate to judicial behavior (e.g., Tetlock et al., 1985), it is also the case that these individual difference variables interact with situational variables (such as being in the majority or minority; Gruenfeld, 1995) to produce judicial outcomes. Neither a strong personality-based nor a strong “situationist” view of judging (or of behavior in general) can be supported by the evidence; rather, interactionism is the proper view (see Funder, 1999, 2001). A salutary by-product of empirical tests of dispositional theories of good judging might well be a greater appreciation for the role of interactions between personality and situational variables in achieving normative compliance.

While I am optimistic that a dispositional approach to evaluating judges will lead to new insights because researchers taking this approach will likely ask different questions than those

30 The point here is not that looking to opinions to find pragmatic judges is wrong-headed; the point is that the character of pragmatism is, at this point, doing no independent work except serving as a label for how this judge ruled. It will be impossible to identify pragmatic judges and predict their rulings if we can only identify pragmatic judges from their rulings. In this approach, pragmatism may be seen as a proper norm for judging judicial acts, but it is now just another of our normative standards rather than part of some character-based theory of good judging. Of course, this may be the sense most advocates of pragmatic judging intend; few scholars advance as explicitly as Solum does a dispositional theory of good judging.
posed in standard norm-based evaluative approaches (or at least employ new and different norms), I am not optimistic that, ultimately, dispositional approaches to good judging will avoid the usual problems encountered in attempts to study the descriptive-normative gap. For if tests of judicial character remain unmoored from actual decisions, critics will eventually dismiss the dispositional claims as irrelevant or invalid. In the end, we evaluate judges to learn how well they are doing their job, however “doing their job” may be framed, and so we must focus on what judges actually do in relation to what we wish they would do.
References


35


