CALIBRATING LEGAL JUDGMENTS

Frederick Schauer and Barbara A. Spellman*

ABSTRACT

In ordinary life, people who assess other people's judgments typically take into account the other judgments of those they are assessing in order to calibrate the judgment presently being assessed. The restaurant and hotel rating website TripAdvisor is exemplary, because it facilitates calibration by providing access to a rater's previous ratings. Such information allows a user to see whether a particular rating comes from a rater who is enthusiastic about every place she patronizes, or instead from someone who is incessantly hard to please. And even when less systematized, as in assessing a letter of recommendation or college transcript, calibration by recourse to the decisional history of those whose judgments are being assessed is ubiquitous. Yet despite the ubiquity and utility of such calibration, the legal system seems perversely to reject it. Appellate courts do not openly adjust their standard of review based on the previous judgments of the judge whose decision they are reviewing, nor do judges in reviewing legislative or administrative decisions, magistrates in evaluating search warrant representations, or jurors in assessing witness perception. In most legal domains, calibration by reference to the prior decisions of the reviewee is invisible, either because it does not exist or because reviewing bodies are unwilling to admit using what they in fact know and employ. Assisted by insights from cognitive psychology and philosophy, this article examines law's aversion to overt calibration and explores what this says about the nature of law and legal decision-making.

1. INTRODUCTION

Legal decision-makers frequently assess the judgments of other legal decision-makers. The Supreme Court evaluates the judgments of federal courts of appeals and state supreme courts, and those courts evaluate the legal and sometimes the factual rulings of trial courts. Courts also routinely determine the legal

* Schauer is David and Mary Harrison Distinguished Professor of Law and Spellman is Professor of Law, both at the University of Virginia. E-mail: schauer@virginia.edu. We are grateful to the participants at presentations of this article at the Duke University School of Law, the University of Pennsylvania School of Law, and the 2015 Congress of the International Association of Legal and Social Philosophy. Comments, references, and information from Matt Adler, Kate Bartlett, Will Baude, George Christie, Ryan Doerfler, Jill Fisch, Josh Fischman, Andrew Hayashi, Craig Konnoth, Maggie Lemos, Paul Mahoney, Stephen Morse, Neil Siegel, Andrew Vollmer, and Ernie Young have been especially helpful.

© The Author 2016. Published by Oxford University Press on behalf of The John M. Olin Center for Law, Economics and Business at Harvard Law School. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com doi:10.1093/law/law010 Advance Access published on November 8, 2016
sufficiency of judgments by Presidents, governors, members of Congress, state legislators, executive department officials, police officers, school principals, and countless employees of administrative agencies. Judges and magistrates in issuing search warrants similarly assess the reliability of the representations made to them by law enforcement officers seeking the warrant. And of course jurors as well as judges evaluate the credibility of the witnesses who testify in court, witnesses who are relaying their past perceptual or evaluative judgments.

When such situations arise in everyday life, the evaluator of someone else's decision often seeks to inform her evaluation by considering the other decisions made by the decision-maker whose decision is now under review. When a friend recommends a restaurant or movie to us, we like to know something about her reactions to other restaurants or movies, preferably including ones we have both patronized, in order to assess whether her judgment on this occasion is worth following. If another friend tells us that we will like someone we have yet to meet, as in the standard blind date scenario, we want to know whom the friend likes and dislikes in order to decide what to make of the endorsement on this occasion. And when an applicant to graduate school reports having an undergraduate GPA of 3.6, the graduate school must assess where a 3.6 stands in relation to other students (hence the value of class rankings), just as a prospective employer desires analogous information, and just as the recipient of a recommendation wants to know whether the recommender tends to like everyone, or no one, or something in between.

In ordinary talk, this process of gauging another's decisions is often described as calibration, although, as explained below, this ordinary sense of calibration diverges from an important technical sense. Indeed, "calibration" may be a covering term encompassing several different processes, which it will become important to distinguish. But for the moment we designate as "calibration" the process by which a user of some measuring device (or an assessor of some measurer) attempts to establish the relationship between the indicated measure and some relevant standard.¹ Thus, we calibrate a (weight) scale by attempting to establish the relationship between what the scale reads and what something actually weighs. If the scale systematically reads five pounds low, and if (perhaps counterfactually) we wish to know what we actually weigh, we calibrate the scale (or calibrate ourselves to the scale) by adding five pounds to the scale's reading or by adjusting the scale to take account of the error. So too when we take a measurement knowing that some ruler gives a measure slightly longer than

---

¹ The Oxford English Dictionary (Oxford 1971, 318) says that to calibrate is to "make allowance for [the] irregularities" of a measuring instrument.
actual length,\(^2\) or when a rifleman aims lower than where the gunsight tells him to aim, knowing from past experience that the gunsight leads him consistently to miss high.

At times this kind of calibration is systematized. Colleges and universities sometimes provide class rankings or other information so that employers and graduate schools can calibrate whether the 3.6 is extraordinarily good or barely above average.\(^3\) And, increasingly, the user of online restaurant and hotel reviews at sites such as TripAdvisor and Yelp can examine the review history of particular reviewers in order to determine whether a reviewer’s rave review should be discounted because this particular reviewer says nice things about everyplace, or whether another reviewer’s brutally negative assessment should similarly be discounted (or, to put it differently, inflated) because that reviewer has something bad to say about every establishment he patronizes.

Yet although such calibration is ubiquitous in ordinary life, and appears at first glance seemingly desirable, the legal system seems reluctant to follow suit, or at least reluctant to admit doing so. A court evaluating an Administrative Law Judge’s denial of Medicare or Social Security benefits would presumably wish to know whether that ALJ is a frequent denier or instead whether this denial should be entitled to great deference because that judge denies benefits so rarely.\(^4\) And the ALJ evaluating the initial grant or denial by an administrative official would presumably also wish to have a particularized—that is, decision-maker specific—decisional history of the official whose judgment she is reviewing. Similarly, appellate judges reviewing a trial judge’s legal or factual ruling

---

2 Of course neither inches nor pounds have a metaphysical reality. Inches, pounds, meters, and kilograms are human-created standards. But although the standard meter, for example, is a creation of human beings, a ruler is nevertheless considered accurate insofar as what it reads approaches the standard measure.

3 Prior to the emphasis on absolute grade point averages prompted by the formula used by the US News and World Report law school ranking system, some law schools would incorporate into their admissions algorithm the previous grade point averages and law school performances of applicants from particular undergraduate colleges and universities. The law school would thus calibrate (or adjust) the undergraduate grade point average of a particular applicant according to the prior performances of prior applicants from that same school.

against a labor union might want to know whether they are assessing the decision of a judge who typically rules for (or against) unions, so as to calibrate their assessment of the trial judge’s ruling on this occasion. In the same way, the Supreme Court’s review of a lower court’s decision to uphold a death sentence might usefully be informed by knowing whether the court under review is generally sympathetic or hostile to capital sentences. And a jury attempting to determine whether a witness who describes a person as intoxicated is accurate or exaggerating would ideally benefit from information about the full range of instances in which that witness has described someone as drunk, or not.

However sensible it might seem for judges, jurors, and appellate courts to wish to calibrate in just this way, and however much the TripAdvisor model seems to embody a larger rationality, the legal system typically avoids such calibration. Or, perhaps more accurately, the legal system appears to avoid admitting that such calibration occurs, even though it likely occurs quite often. Whatever the judges of a US Court of Appeals might actually be thinking in evaluating a decision by a sentencing District Judge to depart upwards from the Federal Sentencing Guidelines, or whatever those appellate judges might actually know from informal research, personal contact, or hallway gossip, it would be inappropriate for the appellate court to say explicitly in an opinion that it was rejecting the upward departure because this judge is known to be an especially tough sentencer or has a history of upward departures. Indeed, it would even be considered inappropriate for an appellate court openly to seek such information, however much the appellate judges might be aware of the information from previous cases or word of mouth. And cross-examination of a witness about the accuracy of previous similar assessments she has made—of drunkenness, say, or that the condition of a road was “poor”—might be

5 Lay testimony about intoxication is the type of inference or “opinion” that is routinely permitted, even though it involves drawing conclusions from more basic observations. E.g., United States v. Denny, 48 Fed. Appx. 732, 737 (10th Cir. 2002); Singletary v. Secretary of HEW, 623 F.2d 217, 219 (2d Cir. 1980).


7 On departures from the Sentencing Guidelines generally, see Gelacak, Nagel, & Johnson (1996). Upward departures are now constrained by, inter alia, United States v. Booker, 543 U.S. 220 (2004), but we need not delve into the details about departures, or the review of them, other than noting that the standard of review is highly deferential. Gall v. United States, 552 U.S. 38, 45-52 (2007) (mandating abuse-of-discretion standard in reviewing departures from sentencing guidelines).

8 Young v. Illinois C.G. Ry., 618 F.2d 332, 337 (5th Cir. 1980) (reversing trial court’s disallowance of lay testimony about “poor” condition of grade crossing).
excluded by rules limiting cross-examination to matters addressed on direct examination.\(^9\)

Our goal in this article is to examine the ways in which just this kind of calibration might be useful in various legal settings, to trace why the legal system seems to be officially hostile to what so many other aspects of life and decision-making have embraced, and to examine the costs and benefits of more and more explicit calibration than now exists in various legal contexts. Our aim is not to make recommendations for reform. We hope simply to expose for analysis and discussion a topic that has seemed unfortunately ignored in the design and evaluation of legal institutions.

2. THE VARIETIES OF CALIBRATION

As previewed above, we have been using “calibration” as a covering term to describe several different processes—different ways in which an assessor might assess someone else’s assessment. Initially, we begin with the basic idea, applicable to all, that calibration is the process of setting or assessing the relationship between some measuring device or measurer and what is to be measured, in order to conform the measurement to some relevant standard. As in the examples of scales, rulers, and gunsights, a measuring device is accurate insofar as what it reports accords with the relevant standard, and a measuring device is well calibrated insofar as its “judgments” over a range of measurements demonstrate a consistent and therefore predictable difference between what the device indicates and what the underlying “truth” actually is. So a scale that reads 150 pounds when the actual weight is 145 is inaccurate by five pounds, but could be considered well calibrated to the extent that the scale is always five pounds high. Even though the scale is inaccurate, its inaccuracy would be “reliable”—that is, consistent—and a five-pound adjustment would be effective in calibrating the scale (or our use of it) insofar as the adjustment uses the consistency of the scale’s inaccuracy to compensate for the inaccuracy. But if the scale was inaccurate yet not in any reliable way, attempts at calibration would be ineffective.

When we move from the judgments of mechanical devices such as scales, rulers, and gunsights to the judgments of human beings, things become more complex. And because there are multiple ways of calibrating of human judgment, we need to disaggregate the multiple ideas encompassed by the covering

---

\(^9\) Fed. R. Evid. Rule 611 (b). This is a topic on which one of us has done considerable experimental research. Spellman & Tenney (2010); Spellman, Tenney, & Scalia (2011); Tenney, Spellman, & MacCoun (2008); Tenney et al. (2007); Tenney, Spellman, & MacCoun (2006).
term "calibration," a task that is especially crucial because the psychological literature on calibration typically addresses an issue related to but importantly different from the type of calibration that is our primary focus here.

2.1. Calibration in Psychology Research
The concept of calibration that dominates the psychological research focuses on the relationship between the degree of confidence a decision-maker has expressed in some judgment and the actual accuracy of that judgment (Bol & Hacker 2012; Hacker, Bol, & Keener 2008; Krug 2007; Luna & Martin-Luengo 2012). Call this confidence-accuracy calibration. Of course accuracy is itself a relation between a judgment and the underlying reality (or "ground truth"), and the research on confidence-accuracy calibration distinguishes two kinds of accuracy. One is absolute accuracy, sometimes described as "calibration" in a narrow sense, and the other is relative accuracy, also known as "resolution" or "discrimination" (Mengelkamp & Bannert 2010; Nietfeld, Enders, & Schraw 2006). The difference between these two is often important, but the basic idea is that decision makers may be more or less confident in their judgments, and those judgments may be more or less accurate (correct), whether relatively or absolutely. Insofar as the decision maker's degree of confidence aligns with the likelihood that the judgment is correct, the decision maker is considered to be well-calibrated (especially in the absolute accuracy sense). And the decision maker is less well-calibrated to the extent that the degree of confidence over- or under-predicts the likely accuracy of the judgment (Bol & Hacker 2010).

Suppose, for example, that someone judges the speed of a passing car to be fifty-five miles per hour, plus or minus five miles per hour. And suppose she is 80 percent sure that her judgment is correct. If it turns out, over multiple trials, that she is in fact correct 80 percent of the time in assessing speeds within this range, we conclude that she is well-calibrated—her confidence level is a reliable predictor of the likelihood of her accuracy. But if she were accurate only 40 percent of the time that she expressed 80 percent confidence, 30 percent of the time she expressed 50 percent confidence, and 50 percent of the time she expressed 30 percent confidence, she would be poorly calibrated. Although she is overconfident on average (i.e. her confidence overstates her accuracy), there is still no way to use her estimates to predict anything of use.

On the other hand, consider someone who is generally overconfident but in a systematic way; that is, who does not show absolute accuracy but shows good relative accuracy. Take the example from Plous (1993, p. 225), "suppose a decision maker were 50 percent accurate when 70 percent confident, 60 percent accurate when 80 percent confident, and 70 percent accurate when 90 percent confident. In such a case confidence would be perfectly correlated with
accuracy, even though the decision maker would be uniformly overconfident by 20 percent.” Even though the decision maker was inaccurate in her assessments, and even though the decision maker was inaccurate in her degree of confidence in her assessments, the uniformity (reliability) of the overconfidence would allow usable calibration of the decision maker’s conclusions.

2.2 How Legal Judgments are Different

Although confidence-accuracy calibration dominates the psychological literature, it has obvious shortcomings when applied to legal contexts. And that is because legal decision makers rarely articulate the degree of confidence they have in their conclusions and often have no access to some “ground truth.”

Thus, although witnesses at trial, especially in response to cross-examination, may express varying degrees of certainty about the facts and observations they are reporting,10 expressions of anything less than complete confidence are largely absent in the context of judicial (and most administrative) judgments. Justice Brandeis captured the phenomenon well when he observed that he ordinarily convinced himself to a lower degree of certainty (fifty-one percent, he said) than that with which he expressed his judgment in writing an opinion.11 And Ronald Dworkin’s well-known “one right answer thesis” (1985, p. 199; 2006, pp. 41–43, 266 nn. 3–5) might be understood in the same way. It is not as if judges actually believe that there is no plausible alternative answer, Dworkin seems to claim, but the phenomenology of judging is such that judges believe their answer to be correct and other answers incorrect, independent of the actual strengths of those beliefs (Bix 2003, pp. 93–95; Posner 2007, p. 10).

Because judges (as well as police officers seeking search warrants and administrative officials making administrative decisions) are typically loath to describe their degree of confidence in an opinion,12 and because they are

---

10 See United States v. Peoples, 748 F.2d 934, 936 (4th Cir. 1984), holding observation evidence admissible even if “phrased in uncertain terms.”

11 Brandeis was comparing himself to Justice Cardozo, who, Brandeis believed, found it necessary to convince himself 100 percent before reaching a judgment or writing an opinion (Rauh et al. 1979, pp. 5, 12). And see Greenhouse (2005, pp. 12, 18), describing a conversation in which Chief Justice Burger advised Justice Blackmun to suppress his uncertainty and write an opinion expressing complete confidence in his judgment.

12 There are two possible exceptions to the statement in the text. One might arise in the context of civil actions against public officials under 42 U.S.C. §1983 (or, for federal officials, Bivens v. Six Unknown Named Agents, 403 U.S. 388 (1971)), where only violations of “clearly established law,” see Wilson v. Layne, 526 U.S. 603 (1999); Harlow v. Fitzgerald, 457 U.S. 800 (1982), can produce liability in the face of a qualified immunity claim. And the other arises in contexts (legal malpractice actions being the most obvious (Meredith & Brennan 2007)), in which questions about the state of the law are questions of fact. Thus a judge in a civil rights action might rule that the state of the law was sufficiently uncertain that there could be no violation of clearly established law, and an expert
especially reluctant to admit to relative low levels of confidence, the principal psychological concept of calibration is of limited value in most legal contexts. It would be nice to know, in theory, whether a given legal judgment was correct or incorrect and how much confidence a judge had that her opinion was correct, thus enabling an observer to determine the degree to which the judge's confidence was calibrated with the likelihood that the judge reached the correct conclusion. But with ground truths rarely accessible (or existing) for such decisions, and with degrees of confidence even more rarely expressed, this sense of calibration, and the one for which there is the greatest amount of psychological research, is of limited value in considering legal judgments.

2.3 Leaving Confidence Behind—Calibration for Accuracy
Because explicit expressions of degrees of confidence are rare in legal and judicial contexts, the more relevant conception of calibration is concerned not with the alignment between confidence and accuracy, but rather with the simpler question of the alignment between an expressed judgment and the actual fact of the matter. This relation between a judgment and the actual truth is ordinarily understood as accuracy, and we can label measuring the relationship between judgments and the ground truth as *accuracy calibration*.

Accuracy calibration is closer to the examples of the scale that reads five pounds high or the ruler whose indication is an eighth of an inch short. So if we substitute a human observer—a witness—for a mechanical scale, we can imagine a human being who, like the "Guess Your Weight" booths at carnivals, estimates the weight of the people she observes. And if the estimate is consistently five pounds over the actual weight, we calibrate her judgments by subtracting five pounds from each of her estimates. The ensuing adjustment then increases the accuracy of the post-calibration determination.

Suppose now that we are dealing with a witness who is testifying at a trial, or a bystander who has witnessed a crime and is reporting what she saw to the police. The witness or bystander reports that the person she saw running out of the bank waving a gun, carrying a sack, and wearing a ski mask appeared to weigh about 200 pounds. If this were a report to the police, the police officer might (in theory, even if rarely in practice) ask the witness if her estimates of weight were usually high, or usually low, or usually close to accurate. And if the estimate of 200 pounds were part of a witness's testimony at a trial, on cross-examination the witness might, again in theory much more than in practice, be asked about the accuracy of her estimates of weight on other occasions.

witness in a legal malpractice action might testify that the state of the law was, for example, probably such-and-such, but not certainly such-and-such.
Alternatively, opposing counsel might, under a liberal interpretation of the rules governing the use of past acts to prove present conduct, offer evidence about previous weight estimates by this witness that had proved to be inaccurate. Under either scenario, the idea would be to calibrate the accuracy of the witness on this occasion by looking at the degree of her accuracy on other occasions. A history of inaccuracy would lead the rational evaluator of the testimony to discount it, and a history of consistent overestimates would lead the rational evaluator to subtract from the witness’s estimate.

If such a decisional history were raised on cross-examination, the inquiry likely will be barred, perhaps under something like Federal Rule of Evidence 611, which typically excludes matters relating to events other than the ones now being litigated. However useful it might be to the trier of fact to calibrate the witness’s testimony in the way just described, the legal system appears to resist allowing a trier of fact to calibrate a factual report by examining the accuracy of other similar reports made by the same observer.

We will return to the factual witness example presently, but even this form of calibration may be of limited relevance to the kinds of legal, as opposed to factual, judgments often made by courts and other legal actors whose judgments are being reviewed. Unlike estimates of weight and other factual reports, locating the ground truth of a legal judgment is more elusive. It is not impossible, of course. A reviewing court might wish to know, for example, how often a trial judge had made obvious errors of law occasioning reversal. The court might engage in rigorous scrutiny of the legal judgments of a trial judge known

---

13 This is a bit of an overstatement, partly because jurisdictions vary with respect to whether they have wide or narrow scope limitations on cross-examination (Mueller & Kirkpatrick 2012, pp. 603–605), partly because so much is left to the trial judge’s discretion, and partly because trial judges vary in how widely they understand Rule 611’s allowance of cross-examination on “matters affecting the witness’s credibility.” As a result, we claim only that the perspective we offer here might justify a broader scope of cross-examination and rebuttal evidence than now generally exists in both the federal and state systems.

14 Expert testimony is an obvious exception, and it is common for cross-examination to focus on the other assessments made by that expert, especially when the cross-examination is directed to determining reliability under Daubert v. Merrill Dow Pharmaceuticals, Inc., 509 U.S. 579 (1993), Kumho Tire Co. v. Carmichael, 526 U.S. 137 (1999), and Fed. R. Ev. 702.

15 Implicit in this statement is the belief that reversal is some (admittedly imperfect) guide to the correctness of a legal judgment. As with John Marshall’s observation in Marbury v. Madison, 1 Cranch (5 U.S.) 137 (1803), that a law allowing a conviction for treason on the testimony of only one witness would be plainly unconstitutional in light of the two-witness rule in Article III, Section 3, there are legal decisions that are simply wrong independently of some court declaring them so (Hart 2012, pp. 124–147). We recognize however, that, especially at the appellate level, and especially in light of the selection effect (Priest & Klein 1984), such examples of plain legal inaccuracy are rare. But insofar as some legal outcomes are right or wrong, frequency of reversal might indicate how often a judge reached the wrong result.
to be frequently reversed for making obvious mistakes of law, while at the same
time being highly deferential, under conditions of legal uncertainty, to the
judgments of a trial judge whose decisions on matters of law were routinely
upheld on appeal. The reviewing court would use the reversal rate to calibrate
the likely accuracy, under conditions of legal uncertainty, of the trial judge’s
legal conclusions.
Although this kind of calibration is possible, in practice, and especially given
the way the selection effect makes cases involving clear answers disproportio-
ally unlikely to be litigated (Priest & Klein 1984), it is rare that decisions on
matters of law, or even of mixed questions of law and fact, can be characterized
as simply right or wrong. Rather, such decisions typically involve questions
about which there is no ground truth or in which we do not know what the
ground truth is. In such cases, a reviewing body may be less concerned with the
degree of accuracy of some reviewed decision as a matter of ground truth than it
is with evaluating the evaluative judgment that is being reviewed. Here the
concern is not with factual or even legal accuracy, but with evaluating—
trying to understand—an evaluation.

2.4 The Calibration of Evaluative Judgments
Although appellate courts and other reviewing institutions sometimes review
factual and legal determinations with relatively clear right or wrong answers, the
assessment of the judgments of others more often arises in contexts in which the
judgments being assessed are more evaluative than factual. The question then
turns to determining how courts assess an assessment, and how they evaluate an
evaluation. When an appellate court evaluates a lower court’s determination
that a defendant had (or did not have) the effective assistance of counsel,\(^\text{16}\) that a
warrantless search was or was not reasonable,\(^\text{17}\) that a state interest was or was
not substantial\(^\text{18}\) or important,\(^\text{19}\) that a regulatory mechanism was or was not
the least restrictive\(^\text{20}\) or that a defendant should or should not prevail on summary
judgment because of the absence or presence of a genuine dispute about a

\(^{16}\) See Glasser v. United States, 315 U.S. 60, 76 (1942), which first used the phrase, although the prin-
ciple was held to be embodied in the Sixth Amendment in Powell v. Alabama, 287 U.S. 45 (1932).

\(^{17}\) U.S. Const. amend. IV (prohibiting “unreasonable searches and seizures”).

interest must be “substantial” to justify restricting commercial speech).

\(^{19}\) See Craig v. Boren, 429 U.S. 190, 197-98 (1976) (discrimination on the basis of gender permissible
only if it serves an “important” governmental objective).

material fact, the appellate court is faced with the task of evaluating what is itself an evaluative judgment. In such cases, one might think that the evaluator would wish to know just what scale the original decision-maker employed in making the decision now under review.

In such review contexts, calibration takes on a different meaning, which we can label evaluative calibration. Just as the graduate school or employer wishes to know what a 3.6 from some university means, and just as the potential hotel or restaurant patron wishes to know what two stars means, 21 so too might evaluators wish to calibrate the kinds of earlier evaluations that have no intrinsic meaning, or at least have range of meaning broad enough that there is little conception of a plainly right or wrong answer. And although one might be (and should be) a scientific or metaphysical realist—a believer in a mind-independent reality—about water and gravity and gold, and even about the rightness of altruism and the wrongness of child abuse, there are few metaphysical realists about the star ratings for hotels and restaurants, the wine ratings on the 100-point scale commonly used by wine experts, and even the idea of an A— or a B+ on a grade scale. And so when we want to know whether an 88-point wine or a two star rating or a 3.6 grade point average is good or mediocre, we would like information about the other ratings of the rater. 22 If the rater has given a restaurant three stars out of a possible four when we thought the restaurant terrible, and if the rater consistently gives high ratings to restaurants we believe on the basis of our own experiences to be mediocre, we might ignore or discount the rater’s rating of an establishment we are considering patronizing for the first time. Similarly, when some law schools algorithmically lowered (or raised) the grade point averages of students coming from particular institutions, the adjustment was based on having seen how students with those grade averages from those schools actually performed in law school. If students from some undergraduate institutions consistently underperformed their undergraduate grade point averages compared to students from other undergraduate


21 Even apart from questions of evaluative calibration, we would also, of course, want to know just what the scale is. For restaurants, for example, three stars is the best you can do in the Michelin Guide, four stars is the maximum for the New York Times, and other guides go up to a maximum of five stars.

22 The Michelin Guide says that three-star restaurants are “worth a special journey.” But if half the restaurants in the Guide were described as being worth a special journey, we might be more reluctant to actually make the journey than if such a rating, as is the case, is given to less than one percent of the establishments rated.
institutions, this differential would be reflected in an adjustment—a calibration—of the admission index.

Thus, when we speak of evaluative calibration, we are not primarily interested in accuracy. Rather, our concern is with how to understand what someone else’s evaluative judgment means in light of the other judgments that that decision-maker has reached, and thus in light of what we can infer that decision maker’s evaluation scale (or biases) to be. If our evaluation of the earlier evaluation takes account of this knowledge, we have engaged in a process of calibration.

3. A NORM OF NONCALIBRATION?

We have offered a few potential applications for legal calibration, and the time has come to explore these and others in somewhat greater depth. We do so in order to hypothesize what appears at first sight to be a norm of noncalibration, the apparent norm of judicial behavior prohibiting or discouraging courts from examining the previous judgments of the body under review in order to calibrate those judgments, or from officially and openly acknowledging such calibration even if it has occurred surreptitiously.

To start with a relatively straightforward example, consider a Supreme Court Justice with no strong preferences one way or another about the death penalty, and who is faced with evaluating a decision by a state supreme court upholding a death sentence against defense claims of, for example, procedural defects, ineffective assistance of counsel, or cruel and unusual administration. That Justice might wish to know whether the state court whose judgment is being reviewed almost always affirms death penalty sentences, or instead almost always vacates them. If the state court had a persistent history of affirming capital sentences against such objections, the reviewing Justice might suppose

23 There is a highly relevant literature on the evaluation of so-called biased reviewers of academic papers (Kuhlisch et al. 2016; Lauw, Lim, & Wang 2008; Roos, Rothe, & Scheuermann 2010), but the assumption of “bias” adds an unnecessary complication. To engage in calibration for the purpose of identifying biased reviewers might in some contexts be useful, but that is merely the extreme subset of the value of calibrating all reviews by all reviewers.

24 In saying that calibration sometimes occurs “surreptitiously,” we imply nothing pernicious, but only suggest that judges often know of the past behavior of the legislatures, courts, and agencies whose judgments they are reviewing, and might well be influenced by that knowledge even as they believe it to be the kind of information they should not take into account. See Wistrich, Guthrie, & Rachlinski (2005), reporting a study in which judges were often unable to ignore potentially relevant information they possessed but knew was legally unusable.

25 See Fried (1992), recounting the persistent anti-death penalty actions of the Ninth Circuit in the early 1990s.
that this case presented a more or less typical case, and would evaluate the
decision below according to the standard she generally applied to such matters.
But if the court being reviewed was one that often or almost always vacated
capital sentences, the reviewing Justice might then calibrate this decision ac-
cordingly, concluding that here the grounds for reversal might be especially
weak. Conversely, if the case arose on a government appeal from a lower ap-
pellate reversal, the reviewing Justice might conclude that a reversal of a capital
sentence by a court that routinely upholds such sentences would rank very high
on the reviewed court’s scale of error. Insofar as the reviewing Justice has
information of this variety about other decisions by the court being reviewed,
she is calibrating the current decision, or, more precisely, is calibrating her
attitude to the current decision in light of what she knows from other cases
to be the relevant scale.

If we posit that actually knowing these other results would enable the re-
viewing Justice to more accurately calibrate the decision under review, or more
precisely calibrate her degree of deference to the court being reviewed, the
question is then whether she should be permitted to examine those other de-
cisions. And if she did calibrate on the basis of other decisions not now before
her, should this practice be openly acknowledged, for example by making ref-
erence to these other decisions in an opinion? Could a judge actually say she is
applying especially close scrutiny on review, for example, because the previous
decisions by the particular court or judge being reviewed invite skepticism?

The death penalty example is unrepresentative in some ways, but represen-
tative in others. It is definitional of the appellate process that judges are

26 The phenomenon here is related to the Supreme Court’s occasional practice of signaling extreme
easiness by assigning the opinion to the Justice least likely on the basis of past performance to be
perceived as sympathetic to the claim now being upheld. Obviously unanimity itself will sometimes
send such a signal, and Brown v. Board of Education, 347 U.S. 483 (1954), Cooper v. Aaron, 358 U.S. 1
(1958), and United States v. Nixon, 418 U.S. 683 (1974), are well-known examples. But a signal is also
being sent when Justice Rehnquist writes for a unanimous Supreme Court in Jenkins v. Georgia,
418 U.S. 153 (1974), making clear the limits of the “local standards” idea in obscenity law, and when
Justice White writes the opinion for a unanimous Court in Sable Communications v. FCC, 492 U.S.
115 (1989), dealing with the limits of obscenity and communications indecency law in the context of
a ban on sexually explicit telephone services. In such cases, assigning the opinion to the Justice
known to be least receptive to claims now being upheld is a way of telling the audience for the
opinion that the case is especially easy, with the signal being dependent on an implicit calibration by
the audience for the opinion.

27 Implicit in the statement in the text is the idea that standards of review or degrees of deference
designate a range, such that there may be more or less deference, or more or less intrusive review,
even within a stated standard. Thus, for example, de novo review may be more or less truly de novo,
and courts reviewing for abuse of discretion may have more or less stringent standards for what
counts as an “abuse.”
reviewing the decisions of other judges, and whether it be a death penalty appeal, a grant or denial of a motion for summary judgment, a decision to support or reject a constitutional challenge to some state law or practice, or any of a large number of other contexts in which there is an appellate review of an evaluative decision below, the basic dynamic remains one of a legal evaluation of an earlier evaluative legal judgment. And especially when the governing law requires that the evaluation be other than de novo, the reviewing judges would seem to benefit from being able to calibrate the judgments they are being asked to review. And those reviewing judges would also seem to benefit by being able to know as much about the other judgments of the reviewed judge as a reader of a restaurant review benefits from knowing about the other judgments of the reviewer. An appellate court reviewing a trial judge’s grant of a motion for summary judgment might be more inclined to defer if it knew that this particular trial judge granted such motions very rarely, and might, conversely, look especially closely if it knew that the judge was one who granted a much greater number of such motions than did her colleagues on the trial bench. Similarly, a magistrate might be skeptical about a search warrant application coming from a police officer with a history of requesting warrants with barely any probable cause, but inclined to issue the warrant when requested by an officer with a history of airtight warrant applications. And an Administrative Law Judge reviewing an agency official’s denial of, say, a Social Security claim might benefit from knowing whether the proclivities of that particular agency official inclined toward granting claims or instead inclined toward denying them.

However sensible such calibration seems, what appears to be a norm of noncalibration seems to preclude obtaining and using knowledge of such other judgments. Congress, state legislatures, administrative officials, administrative law judges, lower court judges, and other individuals and institutions subject to review all make numerous decisions, and thus any particular decision under review lies somewhere on a scale for the reviewee individual or

---

28 Appellate courts often review jury verdicts, but in such cases the appellate court is typically reviewing the decision of a trial judge to permit the case go the jury, or to refuse to set aside a jury verdict.

29 When review is genuinely de novo, the basis for the decision being reviewed might make little or no difference. But if the standard of review is anything above de novo, some degree of deference is required, and it is in those situations where the reviewing body would like, we suppose, to have some idea of where on some scale the particular decision lies for the body being reviewed. When the Supreme Court is engaged in genuine rational basis review, for example, as in cases like New Orleans v. Dukes, 427 U.S. 297 (1976), or Williamson v. Lee Optical Of Oklahoma, Inc., 348 U.S. 483 (1955), it is being extremely deferential to the judgment below, and in evaluating that judgment the appellate court might wish to know just what kinds of decisions that body makes, so as better to be able to calibrate its implicit standard of review to the decision and the decision maker being reviewed on this occasion.
institution. But under a norm of non-calibration, a reviewing court or other body cannot, at least officially and openly, obtain or use this sort of presumably valuable information. Courts cannot admit to looking more carefully in dormant commerce clause cases at legislatures with more of a history of protectionism than other legislatures, or examining more closely the secular legislation of a state with a history of religiously-motivated but seemingly secular legislation, or scrutinizing more closely the motives of states or municipalities known to be especially prone to restricting speech or the press. Appellate courts do not discuss the decisional history of particular trial judges in upholding or reversing grants of summary judgment or motions to dismiss. And whatever magistrates in issuing search warrants actually know or do, they do not admit to requiring stronger showings from some police officers or departments because of what the magistrate knows about the accuracy of previous representations by those officers or departments.

As noted above, similar potential opportunities for calibration also arise in the context of jurors (or judges operating as triers of fact) in evaluating the testimony of witnesses. Here again calibration by reference to the analog of a decisional history seems rarely permitted. Sometimes, of course, witnesses will testify about matters of fact that have straightforward answers or testify about questions where the answer is a simple yes or no. But witnesses also testify about speed, height, weight, temperature, attitude (he seemed angry), condition (he was drunk), and a host of other matters that are as much evaluative (even if not

30 For a state to have a protectionist motive is close to fatal, see, e.g., Bacchus Imports, Ltd. v. Dias, 468 U.S. 263, 269-70 (1984), but sometimes it is not clear whether a protectionist motive underlies a facially nonprotectionist regulation. See, e.g., Kassel v. Consolidated Frightways, Inc., 450 U.S. 662, 676-78 (1981) (suggesting that the state’s purported safety justifications might not have been genuine); Hunt v. Washington State Apple Advertising Comm’n, 432 U.S. 333, 352-53 (1977) (finding North Carolina’s nonprotectionist justifications for its apple inspection program “suspect”). The question is whether a court could, should, or would examine with especially close scrutiny the nonprotectionist justifications offered by a state and an industry with a history of protectionism, as with, for example, New York and the milk industry. See, e.g., H.P. Hood & Sons v. Dumond, 336 U.S. 525 (1949); Baldwin v. G.A.F. Seelig, Inc, 294 U.S. 511 (1935); Mayflower Farms v. Ten Eyck, 297 U.S. 266 (1936).


33 There is some evidence that judges do not take into account previous search warrant outcomes at all (Knight, Gulati, & Levi 2016; Rachlinski, Guthrie, & Wistrich 2011), which suggests that judges do not calibrate based on the decisional history of the requesting officer or department, or anything else.
normatively evaluative) as they are simply factual, and that in important ways can be characterized in terms of a scale. Speed is variable, as is height and weight and so on, but there are also degrees of anger, degrees of drunkenness, and degrees of many other things about which witnesses routinely offer evidence.

When witnesses testify about such scalar matters, the issues are similar to those arising when a reviewing court is evaluating an earlier determination by a lower court, legislature, administrative official, or administrative law judge. Here again the evaluator of a witness’s testimony should want to know where on the witness’s scale some conclusion lies. Indeed, the ideal testimony, in theory, would be testimony in which the assessor—judge or jury—would know about a previous judgment by the witness on some question whose answer is already known by the assessor. If a witness testifies that Susan was angry, it would be ideal, in theory, if the assessor knew how the witness would characterize the degree of anger of Harry, known by the assessor, on an occasion also known by the assessor. This scenario resembles the situation of the restaurant review, for what we would really like to know is how the reviewer rated another restaurant about which we have already formed an opinion. With that information in hand, we would be best able to calibrate the review of the reviewer on this occasion. Analogously, a trier of fact would be best able to calibrate the testimony of the witness on this occasion if the trier of fact could calibrate on the basis of a previous judgment already known by the trier of fact to be accurate or inaccurate, overstated or understated (Spellman & Tenney 2010; Tenney, Spellman, & MacCoun 2006; Tenney et al. 2007).

Such ideal information will, of course, almost never be available. Nevertheless, a second-best solution is for the trier of fact to be able to have information about other judgments made by the witness, even if not about things already known to the trier of fact. But at least if the trier of fact knew about the other judgments, and could form some impression about the alignment of those judgments with some assumed reference standard, the trier of fact could engage in a better calibration of the testimony than might be possible under current practice, where even with respect to witnesses at trial a norm of noncalibration appears to make such matters largely inaccessible.34

34 Witnesses can be cross-examined or impeached on issues going to their credibility. Fed. R. Evid. 607, 611 (b). And cross-examination or impeachment going to witness credibility is typically not limited by the “scope of direct” rule. See, e.g., United States v. Moore, 917 F.2d 215, 222 (6th Cir. 1990). But credibility is rarely understood so broadly as to allow wide-ranging cross-examination or impeachment going to the kind of calibration we are discussing here, especially because credibility is widely understood as focused, even if not exclusively, on issues of veracity and not on issues of perception or judgment.
4. PERHAPS NOT

The hypothesis just offered is that there is a prevailing norm of noncalibration. Under this hypothesis, courts and other reviewing bodies do not typically calibrate their standard of review or degree of deference to what they know or might find out about the previous decisions of the body being reviewed.

It turns out, however, that the hypothesized anti-calibration norm has so many exceptions as to cast doubt on whether such a norm exists at all. We know, for example, that courts sometimes give particular deference to nonbinding precedents from other jurisdictions when the opinion was written by a highly-respected judge (Henry Friendly, for example) (Klein & Morrisroe 1999) or when a judge or a court is thought to have special expertise on some topic (as with, historically, the Second Circuit or the Delaware Supreme Court in business-related cases (Solomine 2005, pp. 1350–1361)). When such deference occurs, the deferring court is in effect aggregating the decisional history of the court or judge to which it is deferring, and thus in effect calibrating the opinion based on other opinions by the relevant authority. Or consider the “cert pool” memos prepared by law clerks and used by Supreme Court Justices in helping them to decide when to grant certiorari. It turns out that such memos bear the name of the clerk who wrote the memo (Wallander & Benesh 2014), thus suggesting that the Justices are calibrating the memo’s recommendation based, at least in part, on their view about the reliability of the clerk who was the author of the memo.

A much more significant counterexample to the hypothesized norm of noncalibration appears to exist with respect to judicial review of administrative agencies. In this context, there is evidence that the past decisions and behavior of an agency being reviewed exert some influence on the attitude of the reviewing court on a particular occasion (Levy & Glicksman 2011). This practice of “agency-specific” standards of review appears to have started with reactions to National Labor Relations Board practices that diverged from those of other

35 See Morin Bldg. Projects Co., v. Baystone Const. Co., 717 F.2d 413 (7th Cir. 1983), in which Judge Posner defers to a judge who is “an experienced Indiana lawyer.” 717 F.2d at 416-17. To the same effect is Judge Easterbrook’s opinion in Covalt v. Carey Canada, Inc., 860 F.2d 1434, 1440 (7th Cir. 1988) (deferring to judges who were “experienced Indiana lawyers”). See also Klein & Morrisroe (1999).

36 This appears to be a one-way phenomenon. We know of no case in which a court has explicitly discounted the weight of some nonbinding authority because of a negative view about some judge’s or some court’s other opinions.

37 Justice Blackmun apparently also wanted to know the name of the judge who decided the case below (Wallander & Benesh 2014), presumably for the purpose of calibrating based on the identity and presumed competence (or not) of the judge or judges below (Cameron, Segal, & Songer 2000).
agencies (Flynn 1995; Winter 1968), in particular the use by the NLRB of adjudication as a rulemaking tool in order to avoid the structures and strictures imposed by the Administrative Procedure Act on agency-rulemaking, and also a differentially (compared to other agencies) large gap between articulated standards and adjudicative outcomes. As a result, it appeared to some observers (although not acknowledged by the reviewing courts) that judicial review of NLRB decisions was more intrusive than review of other agencies, a consequence of the unadmitted judicial knowledge of the differential agency behavior.

More recently, others have both noticed and endorsed the widespread existence of the same phenomenon (Nou 2013) and have endorsed it, with Richard Pildes (2013), for example, arguing that taking differences among agencies and their past behavior into account in evaluating the decisions of those agencies is a sensible reaction to differences among agencies in their political makeup, their structure, the method by which their senior members are appointed, and the matters they are called up to decide. Failing to do so, Pildes argues, represents a formalistic willingness to treat all agencies the same when in fact they are plainly not.

We do not purport here to make a contribution to administrative law. Yet the fact that the past decisions of a particular agency are taken to be relevant to courts in reviewing agency decisions suggests a possible generalization. If it is at times useful and appropriate for reviewing courts to take an agency’s past judgments into account in determining the degree and type of scrutiny to be applied to that agency’s judgment on a particular occasion, then might it also at times be useful and appropriate for a reviewing appellate court to do the same with different trial courts and trial judges? And might it be useful and appropriate for a magistrate to do the same with different police officers and different police departments seeking search warrants, or for an administrative law judge to do the same thing with respect to the different officials and different parts of the agency whose judgments she is asked to review?

It turns out that such history-based calibration practices do in fact exist. The Securities and Exchange Commission, for example, is frequently required to evaluate the accuracy or nonmisleadingness of representations made in registration statements, proxy statements, and periodic reports. But the Commission

38 Unlike Article III federal judges, Administrative Law Judges are “audited,” and their overall decisional record is documented (Mashaw 1978; Mashaw, Merrill, & Shane 2009). There is no evidence, however, of whether such information is taken into account when an Administrative Law Judge’s decision is being reviewed.

for some of its decisions employs an overt process of allowing some registrants the benefit of fast-track or similarly cursory review process, a process available only to registrants with proven track records of accuracy, and a process revocable upon evidence of inaccuracy on some occasion. In evaluating the accuracy of such filings, therefore, the Commission is openly calibrating its degree of scrutiny to what it knows from the past practices of the individual or entity being reviewed. The audit practices of the Internal Revenue Service are similar even if less overt and less (publicly) systematized, where again the degree of scrutiny at the audit and subsequent stages appears to take into account the particular history of the particular taxpayer. And of course the very fact that closer police scrutiny of "the usual suspects" has such social resonance, and not only because of Casablanca, reinforces the view that past practices are often treated as highly relevant in making current official determinations.

The practice of agency-specific review, as well as the other examples noted here, suggests not only that the form of calibration we describe is well known to the law, but also that the hypothesized norm of noncalibration does not in reality exist. In multiple contexts, the evidence suggests that reviewing bodies do appear to calibrate their assessments of the judgment being reviewed in light of the full array of judgments made over time by subject of the review, just as the user of a TripAdvisor review calibrates her assessment of a review in light of the full array of reviews made by a particular reviewer. It is possible that agency-specific review remains more the exception than the rule, but it seems more likely that it is not so much that there is a norm of noncalibration as there is a norm against official acknowledgment of calibration. Even when calibration

---


41 Ironically, the SEC is also one of the agencies whose own judgments appear in an agency-specific way to be subject to especially close scrutiny. “This history of repeated invalidations of SEC rulemakings by the D.C. Circuit suggests some degree of distrust of the SEC’s policymaking judgments.” Jill E. Fisch (2013, p. 704), analyzing Business Roundtable v. SEC, 647 F.3d 1144, 1149 (D.C. Cir. 2011), and, less directly, Business Roundtable v. SEC, 905 F.2d 406, 412 (D.C. Cir. 1990). See also Fisch (2013), p. 705, observing that the D.C. Circuit, “because of its location and specialized caseload [ ] is particularly sensitive to the functioning of [particular] regulatory agencies”).

42 The Internal Revenue Service uses what it calls a Discriminant Function algorithm to identify audit targets, but the details of the algorithm are highly secret Schauer (2003, pp. 159–167). And in a recent listserv posting, a former IRS lawyer observed that “if we audit you and found something, we audited you for the next five years” (Jaffke 2016).

43 And so too in assessing TripAdvisor itself, insofar as we pick guides and guidebooks based on our judgments of the past correlation between the guidebook’s recommendations and our experiences and preferences.
occurs, as in the administrative law context, it appears that the presence of calibration is more a conclusion reached by external observers, as in the NLRB example, than it is something that is openly acknowledged by the reviewing court.

What seems to emerge from the foregoing is a possible exception-ridden norm against calibration, and a more than possible reluctance to acknowledge calibration even when it occurs. The question is then presented about what might explain the reluctance to engage in or to acknowledge a practice that characterizes not only TripAdvisor, but also the way that most people make most of their decisions in most aspects of their lives.

5. BARRIERS TO CALIBRATION

If we are correct in believing, along with much of the nonlegal world and some of the legal world, that calibration is often potentially valuable, if we are correct that calibration will often be assisted by knowing about previous judgments made by the individual or body being reviewed, and if we are correct that such information is sometimes unavailable and almost universally unacknowledged in the legal system, then we might consider why this is so.

One possibility is that the law imagines itself as a pervasively particularistic institution (Solomon 2014, pp. 1146–1148). Whether it be legal scholars urging that matters be decided “one case at a time” (Sunstein 2001), or the fact-specific and particularistic orientation of the common law (Strauss 2010; Wright 2008), or the general impermissibility of using evidence of acts or behavior on other occasions to prove conformity on this occasion, there are important ways in which the legal system, for reasons of particularistic justice (Solum 1994, 2003) or efficiency or a pervasive legal ideology of particularism, is persistently averse to spending too much time dealing with cases or issues other than the one now under consideration. Whether law is in fact essentially particularistic

44 See Blank v. Sullivan & Cromwell, 418 F. Supp. 1 (S.D.N.Y.), where Judge Motley rejected the idea that her own decisional history was relevant to her capacity to sit in an anti-discrimination action.

45 This is the so-called propensity or character rule, embodied in, for example, Rule 404 of the Federal Rules of Evidence.

46 Which one of us has already spent far too much time and ink challenging Schauer (1991, 2003, 2009). The other of us, however, has been more concerned with the frequency with people are often—depending on context—particularists (Spellman 2011).

47 The concern about focusing only on the case under decision seems to underlie not only the limitations on cross-examination in Fed. R. Evid 611, but also the restriction on the use of extrinsic evidence to impeach a witness, as per Fed. R. Evid 608 (b). As to the latter, see the worry about considering matters that are “remote[,] from the incident involved.” Report on Rule 608(b) of the House Committee on the Judiciary, in Fisher (2015, p. 129).
is a difficult question, but it appears as if the law believes that it is true. And whether particularism is desirable is also a difficult question, but, again, the law appears to believe that it is desirable. And thus it may be that the aversion to calibrating by use of decisional history, although most overtly manifested in the kinds of examples we have been discussing, is in fact best explained by a particularistic impulse that is far more pervasive in and about law in general.

Perhaps even more significant is the fact that calibration in the sense we discuss here seems both personal—almost ad hominem—and extra-legal. And to the extent that such a perception exists, the taboo about the explicit use of calibrating information about a decision-maker being reviewed may be part of a larger taboo about using nonlegal inputs generally. Insofar as, for example, “federal judges routinely deny that anything other than traditional legal principles guide their decisions” (Solomine 2005, p. 1358), it should come as little surprise that admitting to calibration is decidedly nonstandard. Judicial opinions, and public legal statements more broadly, typically deny the personal element in the law, and typically are committed to perpetuating both the mystery of the law and the myth of the irrelevance of the so-called nonlegal factors. Indeed, although we have hypothesized a norm of noncalibration, the hard and publicly available evidence for this hypothesis appears thin, and a safer hypothesis would be that there is a norm of nonacknowledgment of calibration, a norm whose existence is clearer, and a norm whose explanation is consistent with what we suggest here.

Other obstacles to history-based calibration may be more pragmatic. It is hardly clear, for example, how judges or other legal decision-makers could obtain the kind of information they might need to engage in the process of calibration. How will they find out about other decisions and about how those other decisions turned out? And even if, in a world of easy electronic access to information, such information was available, would there be some way of making the availability more disciplined and systematic and thus more susceptible to verification and challenge? Thus, there is a serious risk that opening the door to this kind of information would lead to such an expansion of the domain of usable legal information that the disadvantages, even if only logistical and pragmatic, would far outweigh the advantages.

48 This is hardly the place to document the full Legal Realist challenge to the claim that judicial decisions are based only on the factors announced in judicial opinions, but the Realist claim that judicial opinions and other official legal statements are only a small part of what actually determines legal outcomes is exemplified in, for example, Frank (1930), Hutcheson (1929), Llewellyn (1930), and Llewellyn (2011), And see generally Kennedy (1986); Leiter (2007); Schauer (2013); Twining (2012).
Moreover, an aversion to calibration may also embody nonepistemic goals. Just as deference in general (Soper 2005) may often reflect a nonepistemic respect for the decision-making powers of others, the unwillingness of an appellate body to examine the decisional history of those it is reviewing may manifest a form of respect, even if not epistemically justified, for those it is reviewing. An appellate court that is saying that it will look more closely at the decisions of trial judge A than of trial judge B because of trial judge A's history of being reversed is expressing a strong disrespect of judge A, something that the system might wish to avoid even if the disrespect is entirely justified. Similarly, nonepistemic institutional equality goals might militate in favor of treating all agencies or all lower courts as equivalent even if they are not. A Supreme Court that is willing, for example, to openly treat the decisions of some courts of appeals with less deference than it treats others is displaying a lack of respect which, even when epistemically justifiable, may be inconsistent with some of the goals of the legal system. So too, a Supreme Court willing to examine more skeptically the justifications offered by some states on some issues than it looks at the justifications of other states on the same issues may be piercing a veil of equality implicit in federalism, even if the veil, again, rests on shaky epistemic foundations.

Yet it is important to recognize that calibration, even under existing procedures, is not absent. In part, legal decision makers calibrate against their own views, or against an assumed average, or on an assumption that the witness or lower court judge is more or less like them. And we suspect that legal decision makers, especially judges, do exactly what we are suggesting here, but do it in the halls by making use of courthouse gossip, or do it by reading the newspapers, or in other ways obtain information that they are, in theory, not supposed to have, and use information that they are not permitted to acknowledge publicly that they possess in the first place. Insofar as this is the case, and thus insofar as there is far more calibration from other events and other decisions than legal actors are willing or allowed to admit, then it is possible that bringing the entire practice out in the open, making it more systematic, and making it more legitimate, may have some salutary effects. On the other hand, if the costs of calibration are such that the optimal amount of calibration is modest, then a norm of nonacknowledgment of calibration may assist in keeping the degree of actual calibration to appropriate levels.

49 There are exceptions here too, the Voting Rights Act of 1965, Pub. L. 89-110, 79 Stat. 437 (1965), being among the most prominent. Indeed, the Act's willingness to treat some states differently from others because of the history of voting denials in some states might be seen as a rare exercise in official calibration.
6. CONCLUSION

The law hovers ambivalently between generality and particularity. Its generality is exemplified by its heavy reliance on rules, on precedent, and on various principles, maxims, canons, and other vehicles of generality. But the law’s particularity manifests itself in, for example, the very idea of common law method, in the calls to make decisions one case at a time or on the facts of particular controversies, and in the reluctance of the law of evidence to allow past practices to be used as proof of current behavior, however epistemically and probabilistically rational such a course of action may be.

Law’s seeming reluctance to allow reviewers officially and openly to take account of the past decisions of the individuals and institutions they are reviewing reflects this ambivalence. By looking only at the particular decision under review, and not calibrating the posture of review on the basis of a history of decisions, reviewing courts and other reviewing institutions embody the particularism that is a large part of the American legal tradition. But generality is also a substantial part of the legal tradition, both in the USA and elsewhere (Schauer 2003). In exercising review without calibrating in light of the reviewee’s history, reviewing institutions choose a form of particularism not only over generality, but over accuracy as well. In some review environments that may be the right choice to make. But in others it may not be, and it is not implausible to suggest that, at times, law may have at least a small bit to learn from institutions such as TripAdvisor.

References


Roos, Magnus, Jörg Rothe, & Björn Scheuermann. 2011. How to Calibrate the Scores of Biased Reviewers by Quadratic Programming. Proceedings of the


