

Bias, the Brain, and Student Evaluations of Teaching

Deborah J. Merritt*

* John Deaver Drinko/Baker & Hostetler Chair in Law, Moritz College of Law, The Ohio State University. Conversations with many colleagues over the years shaped some of the thoughts in this article. For more immediate assistance, I am indebted to Ruth Colker, Andrew Merritt, Daniel Merritt, and Elaine Shoben. Kristin Harlow provided invaluable research assistance.

Bias, the Brain, and Student Evaluations of Teaching

Abstract

Student evaluations of teaching are a common fixture at American law schools, but they harbor surprising biases. Extensive psychology research demonstrates that these assessments respond overwhelmingly to a professor's appearance and nonverbal behavior; ratings based on just thirty seconds of silent videotape correlate strongly with end-of-semester evaluations. The nonverbal behaviors that influence teaching evaluations are rooted in physiology, culture, and habit, allowing characteristics like race and gender to affect evaluations. The current process of gathering evaluations, moreover, allows social stereotypes to filter students' perceptions, increasing risks of bias. These distortions are inevitable products of the intuitive, "system one" cognitive processes that the present process taps. The cure for these biases requires schools to design new student evaluation systems, such as ones based on facilitated group discussion, that enable more reflective, deliberative judgments. This article draws upon research in cognitive decision making, both to present the compelling case for reforming the current system of evaluating classroom performance and to illuminate the cognitive processes that underlie many facets of the legal system.

Bias, the Brain, and Student Evaluations of Teaching

The complaints are never-ending, voluminous, and contradictory. I talk too loud or not loud enough. I walk too close to people and make them nervous. If I look at students, they are nervous. If I do not look at them they are angry. If I call on them, I am picking on them. If I do not call on them, I have a personal vendetta against them. . . .

When I talk to students in an attempt to ascertain what I do that is so different from the other professors teaching the same section of first-year students, they admit that I do no more in class than their white male professors—my class is no more rigorous, no more intimidating, no more work. In fact, they seem to like the class. . . . Most students appear to like the use of overheads, the introductory and periodic summaries, and question and answer periods. . . .

*The only difference appears to be that I am a Black female*¹

Professors of color have published poignant accounts of harshly negative student evaluations.² The few empirical studies examining instructor race and student ratings confirm that minority faculty receive significantly lower evaluations than their White colleagues.³ The contradictory nature of the student comments on evaluations of minority faculty, the high levels of

¹ Pamela J. Smith, *Teaching the Retrenchment Generation: When Sapphire Meets Socrates at the Intersection of Race, Gender, and Authority*, 6 Wm. & Mary J. Women & L. 53 (1999).

² For descriptions in the legal literature, see, e.g., Barbara L. Bernier, *The Creed According to the Legal Academy: Nihilistic Musings on Pedagogy and Race Relations*, 6 Wash. & Lee R.E.A.L. J. 27, 42-43 (2000); Okianer C. Dark, *Just My 'Magination*, 10 Harv. BlackLetter J. 21 (1993); Richard Delgado & Derrick Bell, *Minority Law Professors' Lives: The Bell-Delgado Survey*, 24 Harv. C.R.-C.L. L. Rev. 349 (1989) (reporting results of survey mailed to all minority law faculty); Tina Grillo, *Tenure and Minority Women Law Professors: Separating the Strands*, 31 U.S.F. L. Rev. 747 (1997); Joyce Hughes, *Different Strokes: The Challenges Facing Black Women Law Professors in Selecting Teaching Methods*, 16 Nat'l Black L.J. 27 (1998-2000); Emma Coleman Jordan, *Images of Black Women in the Legal Academy: An Introduction*, 6 Berkeley Women's L.J. 1 (1990); Reginald L. Robinson, *Teaching from the Margins: Race as a Pedagogical Sub-Text*, 19 W. New Eng. L. Rev. 151 (1997); Pamela J. Smith, *supra* note 1; Donna E. Young, *Two Steps Removed: The Paradox of Diversity Discourse for Women of Color in Law Teaching*, 11 Berkeley Women's L.J. 281 (1996). Cf. Mari J. Matsuda, *Voices of America: Accent, Antidiscrimination Law, and a Jurisprudence for the Last Reconstruction*, 100 Yale L.J. 1329 (1991) (discussing discrimination claims pressed by Asian American professors denied tenure or contract renewal).

³ Until recently, no empirical work probed the relationship between race and student evaluations, a scholarly gap that is itself troubling. Published data on this subject from law schools remain unavailable. A recent analysis of almost 17,000 evaluations completed by undergraduate and graduate students at the University of Texas, however, reveals that minority faculty obtain significantly lower ratings than White professors, even after controlling for tenure status and course type. Daniel S. Hamermesh & Amy Parker, *Beauty in the Classroom: Instructors' Pulchritude and Putative Pedagogical Productivity*, 24 Economics of Educ. Rev. 269 (2005). Another recent study at a predominantly Hispanic university found that both White and Hispanic students evaluated their White professors more highly than Hispanic ones. Jack Glascock & Thomas E. Ruggiero, *The Relationship of Ethnicity and Sex to Professor Credibility at a Culturally Diverse University*, 55 Communication Educ. 197, 204 (2006). Studies have not yet attempted to isolate the reasons for this difference.

express hostility, and the occasional direct references to gender or race raise troubling questions about the role of bias in these assessments.⁴ White faculty members have also noted the possibility of bias in their student evaluations, particularly based on gender, appearance, or political ideology.⁵ Throughout the academy, faculty have questioned whether student evaluations of teaching accurately reflect a professor's success in helping students learn; many charge that evaluations undermine learning by encouraging lenient grading and superficial classroom presentations.⁶ In an increasingly diverse and competitive workplace, can we rely upon conventional teaching evaluations to tell us what we want to know about a professor's classroom success? Or do these evaluations reflect—and perhaps reinforce—biases based on race, sex, and other unwelcome characteristics?

Despite the persistence of these questions, law schools and other academic departments continue to use traditional student evaluations of teaching. Indeed, many professors report that reliance on these measures—particularly on isolated numerical averages—is growing in tenure, promotion, salary, and other decisions.⁷ The academy has been particularly silent in response to

⁴ See, e.g., Delgado & Bell, *supra* note 2, at 385 (referring to “bi-polar” evaluations of minority faculty); Kathryn Pourmand Nordick, *A Critical Look at Student Resistance to Non-Traditional Law School Professors*, 27 W. New Eng. L. Rev. 173, 188 (2005) (student observation that criticisms of Black professors by peers “were often harsh and condescending, while the criticisms of traditional professors were almost backward compliments”); *id.* at 191 (recent review of randomly selected evaluations at a law school yielded “a couple [that] included statements that clearly indicated bias”); Smith, *supra* note 1, at 167 (“The intensity of their anger and hatred [on evaluations] was frightening. Many of them attached notes to their evaluations, espousing crazy racial and/or sexist stereotypes.”).

⁵ See, e.g., Richard L. Abel, *Evaluating Evaluations: How Should Law Schools Judge Teaching?*, 40 J. Legal Educ. 407 (1990); Kathleen S. Bean, *The Gender Gap in the Law School Classroom: Beyond Survival*, 14 Vt. L. Rev. 23, 29 (1989); Martha Chamallas, *The Shadow of Professor Kingsfield: Contemporary Dilemmas Facing Women Law Professors*, 11 Wm. & Mary J. Women & L. 195 (2005); Christine Haight Farley, *Confronting Expectations: Women in the Legal Academy*, 8 Yale J.L. & Feminism 333, 339 (1996); Deborah Maranville, *Classroom Incivilities: Gender, Authenticity and Orthodoxy, and the Limits of Hard Work: Four Lenses for Interpreting a “Failed” Teaching Experience*, 12 Wm. & Mary J. Women & L. 699 (2006).

⁶ See, e.g., Valen E. Johnson, *Grade Inflation: A Crisis in College Education* 237 (2003); Dennis E. Clayson & Mary Jane Sheffet, *Personality and the Student Evaluation of Teaching*, 28 J. Marketing Educ. 149 (2006); Charles R. Emery et al., *Return to Academic Standards: A Critique of Student Evaluations of Teaching Effectiveness*, 11 Quality Assurance in Educ. 37 (2003); Wendy M. Williams & Stephen J. Ceci, “How’m I doing?” 29 Change 13 (Sept./Oct. 1997).

⁷ See, e.g., Judith D. Fischer, *The Use and Effects of Student Ratings in Legal Writing Courses: A Plea for Wholistic Evaluation of Teaching*, 10 Legal Writing 111 (2004); Hugh Hinton, *Reliability and Validity of Student Evaluations:*

questions about racial bias in conventional teaching evaluations. Few articles engage the eloquent critiques that individual minority professors have raised, and schools do not seem to have examined their practices in response to these concerns.⁸

It is time to take seriously the criticisms that scholars have voiced about student evaluations of teaching. Extensive research by psychologists and educators convincingly demonstrates that these evaluations *are* biased. The biases, however, are not simplistic ones based directly on race, gender, or other social categories. Professors do not suffer automatic, consistent penalties premised on race, gender, political ideology, or other commonly recognized categories. Indeed, some “nontraditional” professors obtain very positive teaching evaluations, accurately reflecting their teaching excellence, while some politically moderate, conventional white males receive ratings that seem unduly low.

The bias in student evaluations derives primarily from a relationship that most faculty overlook: the powerful link between student ratings and a small set of nonverbal behaviors.⁹ Conventional student evaluations strongly mirror the professor’s smiles, gestures, and other mannerisms, rather than the professor’s knowledge, clarity, organization, or other qualities more

Testing Models Versus Survey Research Models, 26 PS: Political Sci. & Pol. 562, 562-63 (1993); Richard S. Markovits, *The Professional Assessment of Legal Academics: On the Shift from Evaluator Judgment to Market Evaluations*, 48 J. Legal Educ. 417, 417 (1998).

⁸ See, e.g., Smith, *supra* note 1, at 93-96; Kathryn L. Vaughns, *Women of Color in Law Teaching: Shared Identities, Different Experiences*, 53 J. Legal Educ. 496, 500 (2003) (“One aspect of the law school environment that has especially bothered me is a reluctance to acknowledge that my experiences in the classroom, and those of other people of color, may well be different—sometimes vastly so—than those of my white peers.”). But see Therese A. Huston, *Pedagogy and Social Justice: Race and Gender Bias in Higher Education: Could Faculty Course Evaluations Impede Further Progress Toward Parity?*, 4 Seattle J. Soc. Just. 591 (2006); Nordick, *supra* note 4.

⁹ Although this article focuses on the substantial biases stemming from unconscious reactions to nonverbal behavior, other considerations can also distort law school evaluations. Some studies, for example, suggest that students award lower ratings to faculty if they receive grades before completing evaluations. See, e.g., Dennis E. Clayson et al., *Grades and the Student Evaluation of Instruction: A Test of the Reciprocity Effect*, 5 Academy Mgmt. Learning & Educ. 52 (2006). This dynamic can affect evaluations in courses like legal writing, in which students receive grades throughout the semester. See, e.g., Judith D. Fischer, *How to Improve Student Ratings in Legal Writing Courses: Advice from the Trenches*, 34 U. Balt. L. Rev. 199 (2004); Melissa Marlow-Shafer, *Student Evaluation of Teacher Performance and the “Legal Writing Pathology:” Diagnosis Confirmed*, 5 N.Y. City L. Rev. 115, 115-16 (2002). These problems deserve serious attention, but further discussion of them is beyond the scope of this article.

clearly associated with good teaching. The way in which a professor walks into the room or smiles at the class can affect student ratings much more substantially than what the professor says or writes on the blackboard.¹⁰ Evaluations collected from students after no more than five minutes' exposure to a professor accurately predict assessments gathered at semester's end, leaving little doubt that these evaluations reflect relatively superficial behaviors.¹¹

The nonverbal mannerisms that drive teaching evaluations bear little relation to learning; a professor's ratings on student assessments show little, if any, correlation with objective measures of what students learn in the course.¹² Most of these behaviors, on the other hand, are related to race, gender, and other immutable characteristics; they stem from physiology, culture, and habit.¹³ Social stereotypes, moreover, filter perceptions of these behaviors so that, even when faculty engage in identical classroom behaviors, students may perceive those behaviors differently depending on the professor's race, gender, and other characteristics.¹⁴ Women and minority faculty, therefore, may experience bias on at least two levels. At the same time, white men can also suffer unfairly negative evaluations if they have the "wrong" facial expressions or mannerisms.

These biases do not arise because students are incapable of evaluating teaching; under the right circumstances, the most experienced decision makers will manifest the same biases.¹⁵ The inaccuracies in our current system occur because of the *manner* in which we gather student feedback. Psychology research demonstrates that the human mind functions along two very different tracks, one that generates automatic, instinctive reactions and another that produces more

¹⁰ See *infra* Part I.

¹¹ See Clayson & Sheffet, *supra* note 6 (discussed in detail *infra* notes 64-70 and accompanying text).

¹² See *infra* Part II.C.

¹³ See *infra* Part II.A.

¹⁴ See *infra* Part II.B.

¹⁵ See, e.g., Irene V. Blair et al, *The Influence of Afrocentric Facial Features in Criminal Sentencing*, 15 Psychol. Sci. 674 (2004) (discussing influences on sentencing by trial judges); see also *infra* notes 90-92 and accompanying text (discussing this research).

reflective, deliberative decisions.¹⁶ The way that we currently obtain teaching assessments from students taps their instinctive rather than reflective judgments. Law schools can mitigate the biases in student evaluations by designing evaluation systems that allow students to offer more thoughtful assessments of teaching; the most effective processes would give students additional information about a professor's pedagogic strategies and then engage them in facilitated small-group discussions of teaching.

This first part of this article reviews the psychology research demonstrating the strong link between student evaluations and a professor's nonverbal behavior. The second part examines why this connection is so damaging: It allows race, gender, and other immutable characteristics to bias our assessment of good teaching, while doing little to identify faculty who genuinely enhance student learning. A third section probes the cognitive processes that produce student evaluations, explaining why the evaluation process we use—rather than the students who participate—generates these unreliable results. The final section proposes a system of gathering student feedback, one that makes greater use of facilitated small group discussions, to overcome many of the cognitive biases built into the current system.

Fairness to both students and faculty demands that we look critically at the student evaluation system we currently employ. Understanding the flaws in our teaching evaluation process, moreover, illuminates deeper truths about how the brain works. The legal system depends upon judgments that people make of others, from hiring decisions and due diligence reviews to negotiating strategies and jury deliberations. Appreciating the cognitive channels that distort those assessments will help lawyers improve decision making in all aspects of their

¹⁶ See generally Steven A. Sloman, *Two Systems of Reasoning*, in *Heuristics and Biases: The Psychology of Intuitive Judgment* 379 (Thomas Gilovich et al. eds. 2002); *infra* Part III.

practice. This article begins that process by exploring cognitive paths in judgments that are familiar to all students, faculty, and practicing lawyers: student evaluations of teaching.

I. Nonverbal Behavior and Student Evaluations

Nonverbal behaviors include a wide range of appearances and actions that influence communication apart from verbal content.¹⁷ Smiles, frowns, raised eyebrows, and other facial expressions are nonverbal behaviors. So are shrugs, waves, and other gestures. The ways in which we move, position our bodies, and hold our arms further enrich the nonverbal repertoire. Dress, hairstyle, and other aspects of physical demeanor also contribute to nonverbal communication. So do many “nonsubstantive” aspects of speech, such as voice tone, accent, and cadence.

Humans respond instinctively and rapidly to these nonverbal cues.¹⁸ Nonverbal behavior shapes employment interviews, coaching sessions, and other interactions in the workplace.¹⁹ Expressions, gestures, appearances, and vocal tones influence every aspect of a judicial trial, from jury selection to sentencing.²⁰ Even a simple smile can communicate a wide variety of meanings, from genuine pleasure to discomfort or deception.²¹ We rely constantly on nonverbal signals to detect the emotions, attitudes, and intentions of people around us.

¹⁷ See generally Joseph A. DeVito, *The Interpersonal Communication Book* 180-210 (10th ed. 2004).

¹⁸ See generally Susan V. Choi et al., *The Glimpsed World: Unintended Communication and Unintended Perception*, in *The New Unconscious* 309 (Ran R. Hassin et al. eds. 2005); Bella M. DePaulo, *Nonverbal Behavior and Self-Presentation*, 111 *Psychol. Bull.* 203 (1992).

¹⁹ See, e.g., Ronald E. Riggio, *Business Applications of Nonverbal Communication*, in *Applications of Nonverbal Communication* 119 (Ronald E. Riggio & Robert S. Feldman eds. 2003) (summarizing recent research).

²⁰ See, e.g., Michael Searcy et al., *Communication in the Courtroom and the “Appearance” of Justice*, in *Applications of Nonverbal Communication*, *supra* note 19, at 41.

²¹ See, e.g., Paul Ekman et al., *Smiles When Lying*, 54 *J. Personality & Soc. Psychol.* 414 (1988); Christine R. Harris & Nancy Alvarado, *Facial Expressions, Smile Types, and Self-Report During Humour, Tickle, and Pain*, 19 *Cognition & Emotion* 655 (2005); Julie A. Woodzicka & Marianne LaFrance, *Working on a Smile: Responding to Sexual Provocation in the Workplace*, in *Applications of Nonverbal Communication*, *supra* note 19, at 139.

Despite its cerebral connections, the law school classroom hums with nonverbal behavior. A quick scan of the room will tell the professor which students are paying attention and which ones, more likely, are surfing the web or sending instant messages behind their laptop screens. Even in a large classroom, professors can often sense which students are engaged with the material and which ones are bored, hostile, or confused. Just as much nonverbal communication flows in the opposite direction. From the moment a faculty member walks into the room, students perceive, process, and react to the professor's nonverbal signals.

Researchers have extensively documented the effect of these signals on student evaluations, often contrasting the dominance of classroom "style" over content. One early and well known investigation into these classroom dynamics used a charismatic, distinguished looking, and mellifluous actor to play the role of a scholar named "Dr. Fox."²² The experimenters created a meaningless lecture on "Mathematical Game Theory as Applied to Physician Education," and coached Fox to deliver it "with an excessive use of double talk, neologisms, non sequiturs, and contradictory statements."²³ At the same time, the researchers encouraged Fox to adopt a lively demeanor, convey warmth toward his audience, and intersperse his nonsensical comments with humor. "In short," as one of the investigators summarized, Dr. Fox "gave a very enjoyable lecture in which he offered little or nothing of substance."²⁴

Fox fooled, not just one, but three separate audiences of professionals and graduate students.²⁵ Fifty-five psychiatrists, psychologists, educators, graduate students, and other

²² Donald H. Naftulin et al., *The Doctor Fox Lecture: A Paradigm of Educational Seduction*, 48 J. Med. Educ. 630 (1973).

²³ *Id.* at 631.

²⁴ John E. Ware, Jr. & Reed G. Williams, *The Dr. Fox Effect: A Study of Lecturer Effectiveness and Ratings of Instruction*, 40 J. Med. Educ. 149 (1975).

²⁵ The first group heard Fox's presentation live, while the other two saw a videotape. Naftulin et al., *supra* note 22, at 632-33.

professionals viewed Fox's lecture; not one identified its emptiness.²⁶ On the contrary, Fox received extremely favorable ratings from his audiences. In addition to awarding him strong numerical scores, audience members praised him for an "[e]xcellent presentation," "warm manner," "[g]ood flow," "[l]ively examples," "relaxed manner," and "[g]ood analysis of subject."²⁷

Fox's use of warm, enthusiastic, and lively nonverbal behaviors would have been admirable if it had complemented a substantive presentation; most faculty use stylistic elements to engage student interest and motivate learning. The disturbing feature of the Dr. Fox study, as the experimenters noted, is that Fox's nonverbal behaviors so completely masked a meaningless, jargon-filled, and confused presentation. If style can trump substance so easily, even in the minds of a trained, professional audience, then what role do nonverbal behaviors play in more routine student evaluations?

Several researchers followed up on this question by using the Dr. Fox paradigm to conduct controlled classroom experiments.²⁸ These studies used videos that systematically varied a lecturer's content and nonverbal behaviors to examine their relative effect on student teaching evaluations.²⁹ A meta-analysis of this cluster of investigations concluded that nonverbal behaviors dramatically affected evaluations; an entertaining style increased an instructor's ratings by about

²⁶ *Id.* at 633.

²⁷ *Id.* at 632-33.

²⁸ See, e.g., Ware & Williams, *supra* note 24; Reed G. Williams & John E. Ware, Jr., *An Extended Visit with Dr. Fox: Validity of Student Satisfaction with Instruction Ratings after Repeated Exposures to a Lecturer*, 14 *Am. Educ. Research J.* 449, 453 (1977); see also Herbert W. Marsh & John E. Ware, Jr., *Effects of Expressiveness, Content Coverage, and Incentive on Multidimensional Student Rating Scales: New Interpretations of the Dr. Fox Effect*, 74 *J. Educ. Psychol.* 126 (1982) (reviewing earlier studies).

²⁹ One set of studies, for example, used the same actor who had portrayed Dr. Fox to create six video lectures on the biochemistry of memory. Two of the lectures contained four substantive points (low content), two included 14 points (medium content), and two conveyed 26 points (high content). Within each pair, one lecture employed a highly charismatic style while the other avoided engaging mannerisms. The researchers then randomly assigned undergraduates to view the videos, gathering teaching evaluations of the lecturer afterwards. Ware & Williams, *supra* note 24, at 151.

1.2 points on a 5-point scale.³⁰ Lecturers who provided more content, on the other hand, received “inconsistent and generally much smaller” boosts in their evaluations.³¹

Other studies have isolated some of the specific nonverbal behaviors that generate positive student ratings. Based on a detailed analysis of university classes and student evaluations, Harry Murray determined that a professor’s speech patterns, facial expressions, and humor had the greatest impact on student evaluations.³² More learning-focused behaviors, such as giving concrete examples of concepts, pointing out practical applications, repeating difficult ideas, or providing sample exam questions, failed to correlate significantly with student ratings.³³ While the Fox studies suggested that faculty could reap greater evaluation rewards by focusing on style rather than substance, Murray’s investigation sounded a further disturbing note: Even when concentrating on the stylistic elements of their teaching, faculty can more effectively raise student evaluations by using certain facial expressions than by offering concrete examples or repeating difficult concepts.

A recent case study, centered on the eminent psychologist Stephen Ceci,³⁴ further illustrates the substantial connection between a professor’s nonverbal behaviors and student evaluations of teaching. After participating in a short “teaching skills” workshop conducted by a media consultant, Ceci raised his evaluations in an introductory psychology course from an overall

³⁰ Philip C. Abrami et al., *Educational Seduction*, 52 Rev. Educ. Research 446, 455 (1982).

³¹ *Id.* at 452.

³² Harry G. Murray, *Classroom Teaching Behaviors Related to College Teaching Effectiveness*, in *Using Research to Improve Teaching* 21, 23 (Janet G. Donald & Arthur M. Sullivan eds. 1985) [hereinafter Murray, *Classroom Teaching*]. See also Harry G. Murray, *Low-Inference Classroom Teaching Behaviors and Student Ratings of College Teaching Effectiveness*, 75 J. Educ. Psychol. 138 (1983) (reporting earlier study reaching similar conclusions) [hereinafter Murray, *Low-Inference*]; Harry G. Murray, *Effective Teaching Behaviors in the College Classroom*, 7 Higher Education: Handbook of Theory and Research 135, 148-50 (John C. Smart ed. 1991) [hereinafter Murray, *Effective Teaching*].

³³ Murray, *Classroom Teaching*, *supra* note 32, at 25.

³⁴ Ceci is the Helen L. Carr Professor of Developmental Psychology at Cornell University. He has published extensively on the accuracy of children’s courtroom testimony, as well as other subjects, and recently won the American Psychological Society’s James McKeen Cattell Award for lifetime contributions to an area of critical social importance. See Cornell University, College of Human Ecology, Bio Page for Stephen Ceci, <http://www.human.cornell.edu/che/bio.cfm?netid=sjc9> (last visited Jan. 10, 2007).

score of 3.08 (on a 5-point scale) to 3.92.³⁵ Because he wanted to test the impact of the media training, Ceci carefully used the same syllabus, lecture content, audiovisual materials, assignments, and exams in sections of the course taught immediately before and after the training.³⁶ He altered only his vocal “pitch variability” and the extent of his hand gestures between the two versions of the course. Notably, these small stylistic changes dramatically improved Ceci’s score on *every* aspect of the college’s evaluation form, including items like instructor knowledge, organization, accessibility, textbook quality, fairness in grading, and other qualities that would seem unrelated to vocal pitch or gestures.³⁷

Even just describing a professor’s manner to students can affect their evaluations. In one controlled experiment, a guest lecturer appeared before a large undergraduate class on a day when the regular professor was absent.³⁸ Students received written notes from their regular professor describing the guest as an experienced professor from another university who others considered “industrious, critical, practical, and determined.”³⁹ Half of the notes further described the visitor as “a rather warm person,” while the other half identified him as “a rather cold person.”⁴⁰ Nonverbal behaviors often signal a speaker’s apparent warmth or coldness to an audience; the written descriptions primed the students to view the lecturer’s manner through one of these contrasting lenses.

After a 40-minute informative lecture related to their course material, students evaluated the guest lecturer. Those who had been told that he was “rather warm” rated him as significantly

³⁵ Williams & Ceci, *supra* note 6.

³⁶ *Id.* Ceci and his coauthor took elaborate measures to assure similarities between the semesters. Independent raters, for example, viewed videotaped lectures from both semesters and confirmed that their content was identical. Ceci’s experience with the course, which he had taught for almost 20 years, made the controls feasible.

³⁷ Ceci’s average rating on knowledge improved from 3.61 to 4.05; on organization, from 3.18 to 4.09; on accessibility, from 2.99 to 4.06; on textbook quality, from 2.06 to 2.98; and on fairness, from 3.03 to 3.72. All of these shifts, as well as others on the evaluation form, were statistically significant at $p < .0001$. *Id.*

³⁸ W. Neil Widmeyer & John W. Loy, *When You’re Hot, You’re Hot! Warm-Cold Effects in First Impressions of Persons and Teaching Effectiveness*, 80 *J. Educ. Psychol.* 118 (1988).

³⁹ *Id.*

⁴⁰ *Id.*

more intelligent, interesting, considerate of the class, and knowledgeable about his material than did students who had read that he was “rather cold.”⁴¹ Changing just one word in the lecturer’s written biography was enough to shift student perceptions of his personality, knowledge, and teaching effectiveness.⁴²

The most remarkable evidence of the link between a professor’s nonverbal behavior and student evaluations of teaching, however, comes from recent research into “thin slice” judgments. Harvard psychologists Nalini Ambady and Robert Rosenthal coined the phrase “thin slices” in 1992 to describe brief observations of an individual that generate judgments about that individual’s personality, intentions, and other characteristics.⁴³ Through a meta-analysis of forty-four studies in which subjects observed no more than five minutes of a target’s behavior, Ambady and Rosenthal concluded that these very quick observations “provide a great deal of information” and can trigger detailed predictions about another person’s behavior.⁴⁴ Humans, in other words, make social judgments based on very short observations of other people.

⁴¹ *Id.* Several other measures of personality and teaching effectiveness, including modesty, self assurance, and organization, did not differ significantly between the warm and cold conditions. Widmeyer and Loy suggest that this confirms the relationship between a warm personality and particular personality or teaching characteristics, rather than a general halo effect. *Id.*

⁴² A large number of studies further explore the relationship among instructor personality, nonverbal behaviors, and teaching evaluations. These investigations confirm that an instructor’s personality exerts a large influence on evaluations, largely through his or her nonverbal behaviors. One pair of researchers recently concluded that the impact of personality is so large that evaluations “could most accurately be called a ‘likeability’ scale.” Dennis E. Clayson & Debra A. Haley, *Student Evaluations in Marketing: What Is Actually Being Measured?*, *J. Marketing Educ.* 9, 12-13 (Fall 1990). See also Stephen Erdle et al., *Personality, Classroom Behavior, and Student Ratings of College Teaching Effectiveness: A Path Analysis*, 77 *J. Educ. Psychol.* 394 (1985); Kenneth A. Feldman, *The Perceived Instructional Effectiveness of College Teachers as Related to Their Personality and Attitudinal Characteristics: A Review and Synthesis*, 24 *Research in Higher Educ.* 139, 141-42 (1986); James C. McCroskey et al., *Toward a General Model of Instructional Communication*, 52 *Commun. Q.* 197 (2004); Harry Murray et al., *Teacher Personality Traits and Student Instructional Ratings in Six Types of University Courses*, 82 *J. Educ. Psychol.* 250 (1990); Sally A. Radmacher & David J. Martin, *Identifying Significant Predictors of Student Evaluations of Faculty Through Hierarchical Regression Analysis*, 135 *J. Psychol.* 259 (2001); Barbara R. Sherman & Robert T. Blackburn, *Perceived Characteristics and Teaching of College Faculty*, 67 *J. Educ. Psych.* 124, 130 (1975); Marie Waters et al., *High and Low Faculty Evaluations: Descriptions by Students*, 15 *Teaching of Psychol.* 203 (1988).

⁴³ Nalini Ambady & Robert Rosenthal, *Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis*, 111 *Psychol. Bull.* 256 (1992). See also Nalini Ambady et al., *Toward a Histology of Social Behavior: Judgmental Accuracy from Thin Slices of the Behavioral Stream*, 32 *Advances in Experimental Soc. Psychol.* 201, 203-04 (2000) (defining “thin slices”).

⁴⁴ Ambady & Rosenthal, *supra* note 43, at 267.

Ambady and Rosenthal applied this insight to a detailed exploration of the assessments students offer on teaching evaluations.⁴⁵ They obtained videotapes of thirteen different instructors teaching undergraduate courses at Harvard University. The courses spanned the humanities, social sciences, and natural sciences; the instructors' ratings on end-of-semester evaluations were similarly diverse.⁴⁶ For each instructor, Ambady and Rosenthal abstracted 30 seconds of videotape from a single class session: ten seconds from the first ten minutes of class, ten seconds from the middle of the class, and ten seconds from the last ten minutes.⁴⁷ They played these tapes—without sound—to groups of undergraduates who had never met the featured teacher.

These students rated the instructors on fifteen different qualities, including competence, confidence, professionalism, enthusiasm, optimism, and warmth.⁴⁸ An initial group of raters viewed the entire thirty seconds of silent video for each instructor. Subsequent groups viewed redacted versions of just fifteen or six seconds of tape for each instructor.⁴⁹ Despite their very brief exposure to the instructors, these students produced highly consistent ratings of these qualities. After viewing no more than thirty seconds of an instructor's nonverbal behavior, the students substantially agreed with one another about which instructors were more competent, professional, and possessed other classroom qualities.⁵⁰

⁴⁵ Nalini Ambady & Robert Rosenthal, *Half a Minute: Predicting Teacher Evaluations from Thin Slices of Nonverbal Behavior and Physical Attractiveness*, 64 *J. Personality & Soc. Psychol.* 431 (1993).

⁴⁶ *Id.* at 433. The teachers in Ambady and Rosenthal's study were all graduate teaching assistants, but their work has been replicated with full-time faculty. See *infra* notes 58-63 and accompanying text.

⁴⁷ Ambady & Rosenthal, *supra* note 45, at 433. All clips focused on the teacher alone, without showing students.

⁴⁸ The full list of qualities was: accepting, active, attentive, competent, confident, dominant, empathetic, enthusiastic, honest, likable, not anxious, optimistic, professional, supportive, and warm. *Id.*

⁴⁹ The 15-second tapes used five seconds from each of the three portions of class time, while the six-second tapes used just two seconds from each of those portions. *Id.* Nine undergraduates rated the 30-second segments, eight rated the 15-second clips, and eight judged the six-second versions. *Id.* at 433, 437. The researchers used all female students for the study; some previous research suggests that women are better than men at decoding nonverbal behaviors.

⁵⁰ For the first group of students, who viewed 30-second video clips, reliability of the ratings ranged from a low of .60 (accepting, attentive, and honest) to a high of .89 (active and enthusiastic), with a mean of .72. *Id.* at 433. Reliability was comparable among students who viewed 15-second videos, and just slightly lower for those who saw only six seconds of video. *Id.* at 437-38.

Even more startling, these ratings showed highly significant correlations with end-of-semester evaluations the instructors had received from students enrolled in their courses. For students who viewed thirty seconds of silent videotape, their global rating of the instructor's personality (a sum of their ratings on the fifteen variables) produced a correlation coefficient of .76 ($p < .01$) with end-of-semester evaluations, explaining almost 58% of the variance in those evaluations.⁵¹ For students who viewed only *six* seconds of the same taped behavior, the correlation with end-of-semester evaluations was almost as high ($r = .71, p < .01$).⁵² These correlations are strikingly high for social phenomena.⁵³

Ambady and Rosenthal concluded that their findings were “quite remarkable: On the basis of observations of video clips just half a minute in length, complete strangers were able to predict quite accurately the ratings of teachers by students who had interacted with them over the course of a whole semester.”⁵⁴ Even judges who watched as little as six seconds of an instructor's silent

⁵¹ *Id.* at 434. A correlation coefficient expresses the strength and direction of a linear relationship between two variables. These coefficients range from -1.00, which signals a perfect negative relationship between the two variables, to 1.00, indicating a perfect positive association between the two. Most social scientists consider correlations between .1 and .3 (or -.1 and -.3) to be small; those between .3 and .5 to be moderate; and those over .5 to be large. Jacob Cohen, *Statistical Power Analysis for the Behavioral Sciences* 79-81 (2d ed. 1988); Will G. Hopkins, *A New View of Statistics*, “A Scale of Magnitudes for Effect Statistics,” <http://www.sportsci.org/resource/stats/effectmag.html#cohen> (accessed Jan. 3, 2007).

Although correlation coefficients are widely used, their relative sizes can mislead nonstatisticians; a correlation coefficient of .4 is not twice as large as one of .2; it is four times as strong. To correct for this and offer a commonsense way of thinking about correlations, social scientists often refer to the “amount of variance” that one variable explains in another. This amount is the square of the correlation coefficient. A correlation coefficient of .2, for example, means that knowledge of one variable allows us to explain about 4% of the variation in the other variable. A coefficient of .4 means that one variable explains about 16% of the variance in the other. *See generally* Jack Levin & James Alan Fox, *Elementary Statistics in Social Research* 369 (10th ed. 2006).

The designation “ $p < .01$ ” indicates the statistical significance of a correlation or other statistical relationship. The “ p ” level expresses the likelihood that a given relationship would have occurred purely by chance. When $p < .01$, the reported relationship would happen randomly—rather than because of an actual relationship—less than one time out of 100. By convention, social scientists accept p levels of less than .05 as “statistically significant.” *See, e.g.,* Levin & Fox, *supra*, at 230.

⁵² Ambady & Rosenthal, *supra* note 45, at 438. The students who rated the 15-second clips achieved a smaller, statistically nonsignificant correlation with end-of-semester evaluations ($r = .44, p > .05$). *Id.*

⁵³ *See, e.g.,* Cohen, *supra* note 51, at 81 (in social psychology, correlations of .50 are “about as high as they come”); Hopkins, *supra* note 51 (correlation coefficients over .7 are “very large, very high, huge”).

⁵⁴ Ambady & Rosenthal, *supra* note 45, at 435.

classroom behavior “predicted with surprising accuracy” students’ end-of-semester evaluations.⁵⁵ These findings confirmed “the considerable influence of very subtle affective nonverbal behaviors on the teaching process.”⁵⁶ A substantial component of student evaluations, in other words, seems to depend on the professor’s appearance, expressions, gestures, and other nonverbal behavior, rather than on the content of what the professor says.⁵⁷

Elisha Babad and Dinah Avni-Babad later collaborated with Rosenthal to build on the initial Ambady-Rosenthal study.⁵⁸ This second study used video clips of forty-seven different professors teaching sixty-seven different courses.⁵⁹ The instructors in this study were experienced faculty, rather than the graduate teaching assistants taped in the Ambady and Rosenthal study; their teaching also spanned a greater range of course types.⁶⁰ Babad and colleagues also were able to incorporate nonverbal aspects of speech in their study by using raters unfamiliar with the language spoken by the professors; thirty-nine American undergraduates, graduate students, professors, or other professionals rated videotapes of professors teaching in Hebrew at the Hebrew University of Jerusalem.⁶¹

Each of these raters viewed nine seconds of tape drawn from each of the sixty-seven classes and evaluated each professor on three scales: “Likable, Warm, Friendly,” “Competent,

⁵⁵ *Id.* at 438.

⁵⁶ *Id.* at 440. The correlation between thin slices of behavior and teaching evaluations is so large, in fact, that these correlations are the highest researchers have obtained in thin-slice research. Ambady et al., *supra* note 43, at 217-20.

⁵⁷ An earlier study, conducted with university accounting instructors, reached very similar results. Ten judges viewed 4-minute clips of teaching behavior by each of nine instructors. The judges agreed on how these nine instructors should rank on an overall measure of effectiveness. Their rankings, moreover, correlated significantly with the instructors’ rankings on end-of-semester evaluations ($r = .70$). Ambady et al., *supra* note 43, at 209 (reporting results of study by Stallings and Spencer).

⁵⁸ Elisha Babad et al., *Prediction of Students’ Evaluations from Brief Instances of Professors’ Nonverbal Behavior in Defined Instructional Situations*, 7 Soc. Psychol. Educ. 3 (2004).

⁵⁹ *Id.* at 9.

⁶⁰ *Id.* at 7.

⁶¹ *Id.* at 8-10. “Nonverbal” aspects of speech include pitch, pauses, speed, and other factors that do not depend on the content of speech.

Effective, Professional,” and “Boring, Passive versus Interesting, Active.”⁶² The researchers then correlated these ratings with end-of-semester evaluations offered by students enrolled in each course. Ratings of the professors’ competence and ability to interest an audience, as manifested in the nine-second video clips, correlated strongly with the enrolled students’ ratings of the professor’s humor, enthusiasm, clarity, and overall classroom performance.⁶³

Most recently, Dennis Clayson and Mary Jane Sheffet reproduced Ambady and Rosenthal’s findings with more than 700 junior and senior college students enrolled in business courses.⁶⁴ Clayson and Sheffet visited fourteen different sections of these courses just after the first class had convened and the primary instructor had introduced him or herself to the students.⁶⁵ The instructor had not yet distributed a syllabus or other materials to the class, and had spoken with the students for less than five minutes. The instructor left the room after this brief exposure, while Clayson and Sheffet asked the students to complete questionnaires about the instructor’s personality, rating the faculty member’s agreeableness, creativity, conscientiousness, stability, and extroversion.⁶⁶ Consistent with Ambady and Rosenthal’s research, these personality ratings—

⁶² *Id.* at 11. The judges showed a high degree of consistency in their ratings. *Id.* at 13.

⁶³ Babad and colleagues combined scores on these evaluation questions into an “instructional” composite score. *Id.* at 15. This score, in turn, correlated highly with ratings for competence and interest drawn from lecture video clips. *Id.* at 17. In this study, the ratings based on brief video clips did not correlate significantly with some end-of-semester ratings, such as those related to course workload and difficulty, instructor accessibility, readings, and course content overall. *Id.* at 15. The failure of these correlations to reach significance is understandable, given their more attenuated relation to characteristics discernible from brief video clips. The difference among ratings in this study, however, does not mean that educators can rely upon some portions of student evaluations to escape the “thin slice” effect. Multiple studies have shown that item-scores on teaching evaluations are heavily intercorrelated. No matter what specific questions a form asks, evaluations appear to tap a student’s global evaluation of instruction. *See, e.g.,* Sylvia d’Apollonia & Philip C. Abrami, *Navigating Student Ratings of Instruction*, 52 *Am. Psychol.* 1198, 1199-1201 (1997).

⁶⁴ Clayson & Sheffet, *supra* note 6.

⁶⁵ *Id.* The fourteen sections included six in Organizational Management and eight in Principles of Marketing. *Id.* at 151.

⁶⁶ These five characteristics, as Clayson and Sheffet point out, are the “big five” personality dimensions. Most psychologists agree that these five dimensions define a major portion of each individual’s personality, are relatively fixed within each person, depend substantially on genetics, and have little cultural component. *Id.* at 151. Individuals high in agreeableness “tend to be friendly, trusting, and cooperative,” while those who are conscientious are “methodical, well organized, and respectful of their duties.” *Id.* Stability denotes people who are “relaxed, less emotional, and less prone to distress,” and creativity characterizes those who are “open minded, creative, and

based on less than five minutes of classroom contact—correlated significantly with conventional end-of-semester evaluations of the instructor’s teaching.⁶⁷ Ratings of agreeability, creativity, conscientiousness, and stability each showed significant correlations with end-of-semester ratings, as did a global measure of personality combining all five dimensions.⁶⁸

Students, in other words, rapidly form an impression of a professor’s personality; that image, based almost entirely on nonverbal behavior, jells within the first few minutes of the semester. Nonverbal mannerisms and other conduct may refine the students’ impression as the semester progresses, but the initial image remains telling. The significant correlation between assessments completed after just five minutes of class and those offered at semester’s end is daunting. Based on this correlation, Clayson and Sheffet conclude that traditional teaching evaluations “follow a seriously flawed paradigm.”⁶⁹ They recommend the initiation of “research and discussions . . . to replace the current . . . system with some other form of evaluation.”⁷⁰

It is tempting to believe that law students are too sophisticated or well educated to react as strongly as other students to a professor’s nonverbal behaviors. There is no evidence, however, to exempt any group of adults from this phenomenon. The subjects who applauded the meaningless

interested in culture.” *Id.* Extroverted individuals, finally, “will seek out the company of others and be energized by such interactions.” *Id.*

⁶⁷ Students assessed the five personality factors by rating their instructor on five 7-point Likert scales. E.g., students indicated whether they found their instructor “disagreeable” or “agreeable,” with 1 indicating most disagreeable and 7 denoting most agreeable. *Id.* at 152. In addition to examining each of the five personality dimensions individually, Clayson and Sheffet averaged the five scores to obtain a “global” measure of personality. *Id.*

The end-of-semester evaluations included six fairly standard rating scales: (1) The instructor created an atmosphere conducive of learning. (2) The instructor explains material appropriately. (3) The instructor shows interest in student learning. (4) The instructor sets high but reasonable standards. (5) Rate your satisfaction with your learning in this course. (6) What grade would you give your instructor? Students answered each of these items with a letter grade (A-F). Clayson and Sheffet averaged the responses to obtain an overall assessment. *Id.* at 152.

⁶⁸ *Id.* at 154. Several other studies have reached similar results, showing very high correlations between evaluations gathered early in the semester and those collected at semester’s end. See Richard G. Kohlan, *A Comparison of Faculty Evaluations Early and Late in the Course*, 44 J. Higher Educ. 587 (1973) (evaluations collected after first class hour correlated highly with those gathered during the last week of the semester); Matthew H. Sauber & R. Rodman Ludlow, *Student Evaluations Stability in Marketing: The Importance of Early Class Meetings*, 88 J. Midwest Marketing 41 (1988) (demonstrating high correlation between evaluations collected during the second week of the semester and those gathered at semester’s end).

⁶⁹ Clayson & Sheffet, *supra* note , at 159.

⁷⁰ *Id.*

lecture delivered in the initial Dr. Fox experiment were all graduate students or professionals, many with M.D.s and other advanced degrees.⁷¹ Similarly, the students whose “five minute” ratings of their professors accurately predicted end-of-semester evaluations were advanced business students with professional ambitions.⁷² Even experienced trial judges react unconsciously to nonverbal behavior when sentencing defendants.⁷³ Responding to nonverbal conduct is not a sign of immaturity, low educational attainment, or carelessness; these reactions are an essential element of how the brain functions.

II. The Shadow Side of Nonverbal Behavior

A connection between nonverbal behaviors and teaching evaluations is not itself surprising. Teaching requires effective communication, and communication entails more than simply uttering words. An enthusiastic, expressive style maintains audience attention, emphasizes main points, and kindles deeper interest in the subject. Stylistic techniques can also enhance clarity and understanding. The fact that a professor’s smiles, emphatic gestures, eye contact, changes in vocal pitch, and relaxed but confident movements correlate with more positive student evaluations of teaching is neither surprising nor threatening. Nonverbal behaviors surely play some role in good teaching, and many faculty members work to polish their classroom style, as well as their substantive knowledge.⁷⁴

The research on student evaluations is troubling, not because it confirms *some* connection between a professor’s style and student evaluations, but because it reveals an *overwhelming* link

⁷¹ See *supra* notes 22-27 and accompanying text. The authors of that study commented specifically on the “educational sophistication” of the audiences they had deceived. Naftulin et al., *supra* note 22, at 633.

⁷² See *supra* notes 64-70 and accompanying text.

⁷³ See *infra* notes 90-92 and accompanying text.

⁷⁴ Cf. Nira Hativa, *Teaching Large Law Classes Well: An Outsider's View*, 50 J. Legal Educ. 95 (2000) (offering advice to legal educators on ways to improve teaching, including tips on nonverbal behavior).

between those two factors: Nonverbal behaviors appear to matter much more than anything else in student ratings. Enthusiastic gestures and vocal tones can mask gobbledygook;⁷⁵ smiles count more than sample exam questions;⁷⁶ and impressions formed in thirty seconds accurately foretell end-of-semester evaluations.⁷⁷ The strong connection between a few nonverbal behaviors and student evaluations creates a very narrow definition of good teaching. By relying on the current student evaluation system, law schools implicitly endorse an inflexible, largely stylistic, and homogeneous description of good teaching. Rather than encouraging faculty to use nonverbal behaviors to complement excellent classroom content, organization, and explanations, the present evaluation system largely eliminates the “dog” of substance, leaving only the “tail” of style to designate good teaching.⁷⁸

Neither law students nor faculty benefit from such a narrow definition of good teaching. The psychology literature, moreover, identifies three further difficulties with the disproportionate role that nonverbal behaviors play in student evaluations. First, the behaviors that most influence these evaluations are rooted in physiology, culture, personality, and habit. Those behaviors are difficult for any faculty member to alter, and they often reflect characteristics like race, gender, nationality, or socioeconomic class. Second, the current evaluation process allows social stereotypes to filter students’ perceptions of instructor behaviors; students see the nonverbal behaviors of some faculty differently than they view the identical behaviors in other professors. These stereotypes further disadvantage faculty based on categories like race or gender. Finally,

⁷⁵ See *supra* notes 22-27 and accompanying text.

⁷⁶ See *supra* notes 32-33 and accompanying text.

⁷⁷ See *supra* notes 45-57 and accompanying text.

⁷⁸ Cf. Clayson & Sheffert, *supra* note 6, at 158 (describing a finance professor who raised his students’ scores on a nationally scaled exam from the 13th to the 97th percentile, but who “consistently placed in the lowest third of all faculty” on student evaluations; “if good teaching is . . . what a student evaluation says it is, then this professor probably should be replaced”); Audesh K. Paswan & Joyce A. Young, *Student Evaluation of Instructor: A Nomological Investigation Using Structural Equation Modeling*, 24 *J. Marketing Educ.* 193, 200 (2002) (“Should instructors be clones when it comes to behavioral manifestations? . . . [If these questions are] ignored, a hallmark of higher educational institutions, that is, diversity in its broadest sense, could become the victim.”).

the ratings that students award through the present evaluation system bear little relationship to objective measures of learning. The current system of student evaluations, in other words, rewards faculty according to relatively trivial indicia of who they are, or seem to be, rather than what they do or accomplish in the classroom.⁷⁹

A. *Nonverbal Behaviors and Mutability*

Some of the nonverbal behaviors that affect student ratings are changeable. Instructors can learn to move around the classroom with more ease; to speak directly to students rather than lecturing from notes; and to gesture more emphatically. With training and practice, some faculty members can improve their evaluations—and their students’ learning—by mastering these kinds of actions.⁸⁰

Many of the behaviors that substantially shape student evaluations, however, are gestures, expressions, tones of voice, and other characteristics that stem from an instructor’s physiology, culture, habit, and personality. These aspects of classroom behavior are “unintended,” “unconscious,” and largely immutable.⁸¹ Professors who manifest these behaviors or appearances will never raise their evaluations beyond a pre-set ceiling, no matter how diligently they work at effective and engaging classroom presentations.⁸²

⁷⁹ These flaws in the evaluation system, of course, do not mean that faculty members who receive high evaluations are bad teachers. A better assessment process, more focused on student learning, might identify many of the same teachers as effective. Eliminating arbitrariness and bias in the system, however, is essential for those who are unfairly penalized.

⁸⁰ The psychologist Stephen Ceci, for example, improved his teaching evaluations dramatically through media training; he recounted his experience in the case study discussed above. *See supra* notes 34-37 and accompanying text.

⁸¹ Ambady & Rosenthal, *supra* note 43, at 256.

⁸² *See* Ambady et al., *supra* note 43, at 205; Clayson & Sheffet, *supra* note 6, at 158 (“if . . . student perceptions are even marginally related to relatively long-lasting traits [in instructors], it may be true that some teachers never will receive consistently high evaluations in certain environments, irrespective of anything they do or possibly could do”); Murray, *Effective Teaching*, *supra* note 32, at 162.

Individuals with symmetric facial features, for example, appear more agreeable, more conscientious, and less neurotic than individuals with asymmetric features.⁸³ Professors with subtle facial asymmetries will seem more worried, nervous, careless, and disorganized to their students, as well as less helpful or sympathetic than their colleagues with more pleasing visages.⁸⁴ Overcoming these negative impressions is difficult, given the rapidity and lasting nature of “thin slice” judgments. Indeed, some research confirms that professors with attractive faces receive more positive student evaluations than those with less desirable features.⁸⁵ Notably, the effect may be stronger for men than women, although it affects both genders.⁸⁶

Similarly, some faces appear more competent than others; some facial structures look immature or unintelligent.⁸⁷ Although researchers have not tested this effect directly in the classroom, they have identified significant effects in several other contexts. Between 2000 and 2004, for example, the candidate with the more “competent” face won more than two-thirds of

⁸³ Fahim Noor & David C. Evans, *The Effect of Facial Symmetry on Perceptions of Personality and Attractiveness*, 37 *J. Research Personality* 339 (2003); Todd K. Shackelford & Randy J. Larsen, *Facial Symmetry as an Indicator of Psychological, Emotional, and Physiological Distress*, 72 *J. Personality & Soc. Psychol.* 456 (1997). The perception appears misplaced; there is little evidence that facial symmetry actually corresponds with personality traits. *See, e.g.*, Bernard Fink et al., *Facial Symmetry and the ‘Big Five’ Personality Factors*, 39 *Personality & Individual Differences* 523 (2005).

⁸⁴ Noor & Evans, *supra* note 83, at 346.

⁸⁵ *See, e.g.*, Hamermesh & Parker, *supra* note 3 (regression analysis of 16,957 student evaluations completed at the University of Texas showed that attractive professors, as rated by undergraduates unfamiliar with the faculty member, received significantly higher evaluations than unattractive ones; the rating difference between most and least attractive faculty constituted a full point on a 5-point teaching evaluation scale); Todd C. Riniolo et al., *Hot or Not: Do Professors Perceived as Physically Attractive Receive Higher Student Evaluations?*, 133 *J. General Psychol.* 19 (2006) (naturalistic study based on www.ratemyprofessors.com, confirming higher ratings for professors perceived as attractive).

⁸⁶ Hamermesh & Parker, *supra* note 3, at 372-73.

⁸⁷ *See, e.g.*, Alexander Todorov et al., *Inferences of Competence from Faces Predict Election Outcomes*, 308 *Science* 1623 (2005); Leslie A. Zebrowitz et al., *Trait Impressions as Overgeneralized Responses to Adaptively Significant Facial Qualities: Evidence from Connectionist Modeling*, 7 *Personality & Soc. Psychol. Rev.* 194, 194 (2003).

contested congressional elections.⁸⁸ Students are likely to incorporate the same biases, attributing more knowledge to faculty with “competent” faces and more warmth to those with babyish ones.⁸⁹

High-quality research likewise demonstrates that individuals with Afrocentric features appear more aggressive than people without those features.⁹⁰ Trial judges respond to that perception by imposing longer sentences on criminal defendants with Afrocentric features than on those without those characteristics: A recent large-scale analysis identified a significant relationship between Afrocentric features and sentence length, even after carefully controlling for severity of the primary offense, concurrent offenses, and prior offenses.⁹¹ The disparity, which affected both White and Black defendants, was substantial; individuals with heavily Afrocentric features received sentences seven to eight months longer than those with identical criminal histories but few Afrocentric features.⁹² The same facial features that influence experienced trial judges very likely affect law students as well, prompting them to view professors with Afrocentric features as more hostile than their other professors.

⁸⁸ Todorov et al., *supra* note 87, at 1624. Competent appearance was judged by individuals from other states who did not recognize the candidates or realize that they were competing for an electoral position. After briefly viewing a single black-and-white photo of each candidate, they indicated which individual they believed was more competent. Competence ratings correlated with election results between 66.0% and 73.3% of the time.

⁸⁹ The overall assessment of faculty with competent or babyish faces may depend on other attributes as well as the particular classroom context. Under some circumstances students may value highly the intelligence of a professor; under others, they may be more moved by a professor’s warmth and support. *Cf.* Todorov et al., *supra* note 87 (noting that, although competent-appearing political candidates win elections significantly more often than those with less mature faces, the latter may secure an advantage in races in which integrity is a primary issue). The interplay of appearance and nonverbal behaviors is complex.

⁹⁰ Irene V. Blair et al., *The Use of Afrocentric Features as Cues for Judgment in the Presence of Diagnostic Information*, 35 *European J. Soc. Psychol.* 59 (2005). *Cf.* Joni Hersch, *Profiling the New Immigrant Worker: The Effects of Skin Color and Height* (Aug. 18, 2006), <http://ssrn.com/abstract=927038> (among recent legal immigrants to the United States, lighter skin color correlated with higher wages).

⁹¹ Irene V. Blair et al., *The Influence of Afrocentric Facial Features in Criminal Sentencing*, 15 *Psychol. Sci.* 674 (2004).

⁹² *Id.* at 677-78. More precisely, a hypothetical White or Black defendant with Afrocentric features one standard deviation above the mean for their race group, and with mean scores for each criminal history variable, would have received a sentence 7-8 months longer than a defendant with the same mean criminal history scores but Afrocentric features scoring one standard deviation below the mean for their group. *Id.* See also William T. Pizzi et al., *Discrimination in Sentencing on the Basis of Afrocentric Features*, 10 *Mich. J. Race & L.* 7 (2005) (summarizing the work of Blair and her colleagues in this area).

Voice quality also affects interpersonal judgments. Individuals with attractive voices seem more competent, powerful, and warm than those with less desirable vocal qualities.⁹³ People who speak in babyish tones sound warmer than other people, but less expert or commanding.⁹⁴ Loud voices register as more authoritative and knowledgeable than soft ones.⁹⁵ A professor's natural vocal tones, therefore, influence student perceptions of the professor's competence, warmth, knowledge, power, and other qualities.

In addition to these physiological features, a faculty member's learned mannerisms significantly affect student perceptions. By adulthood, these characteristics are as much part of us as our physical features. Speech patterns, for example, differ by culture and region. Americans tend to associate rapid speech with competence,⁹⁶ even though some Americans use more leisurely speech patterns. Students in U.S. law schools are likely to prefer fast-speaking faculty members, rating them as more intelligent and knowledgeable than slower speaking professors, even if the two groups of faculty deliver comparable content.

Similarly, White Americans engage in frequent eye contact, believing that it demonstrates honesty, integrity, and attention.⁹⁷ African Americans value eye contact less, and employ it less

⁹³ See, e.g., Diane S. Berry, *Vocal Types and Stereotypes: Joint Effects of Vocal Attractiveness and Vocal Maturity on Person Perception*, 16 J. Nonverbal Behavior 41 (1992) [hereinafter Berry, *Vocal Types*]. "Attractive" voices differ by sex, but people show a high degree of consensus on which male and female voices are most attractive. See Diane S. Berry, *Vocal Attractiveness and Vocal Babyishness: Effects on Stranger, Self, and Friend Impressions*, 14 J. Nonverbal Behavior 141 (1990).

⁹⁴ Berry, *Vocal Types*, *supra* note 93, at 51. Attractiveness and babyishness are separate dimensions of vocal quality, so these characteristics can interact to form a variety of distinct impressions. Attractive and mature voices, for example, appear competent and powerful, but less warm than attractive, babyish voices. Audiences perceive the latter as especially warm, but less powerful and competent. *Id.*

⁹⁵ See, e.g., Ying Peng et al., *The Impact of Cultural Background and Cross-Cultural Experience on Impressions of American and Korean Male Speakers*, 24 J. Cross-Cultural Psychol. 203 (1993).

⁹⁶ *Id.*; George B. Ray, *Vocally Cued Personality Prototypes: An Implicit Personality Theory Approach*, 53 Commun. Monographs 266 (1986). Speakers from other cultures, such as Korea, either disregard vocal speed in judging competence or rate slower speakers as more competent. Peng et al., *supra* note 95, at 215 (native Koreans did not associate vocal rate with competence); *id.* at 215-16 (reporting unpublished results of Lee and Boster, who found that Koreans perceived slower speakers as more competent).

⁹⁷ See, e.g., Elisha Babad, *Nonverbal Behavior in Education*, in *The New Handbook of Methods in Nonverbal Behavioral Research* 283, 290 (Jinna A. Harrigan et al. eds. 2005).

frequently while speaking.⁹⁸ When mulling the answer to a question, individuals in some cultures look up while members of other cultures look down.⁹⁹ Cultural differences like these can prompt a classroom of predominantly White American students to believe that faculty of color or foreign born professors are less attentive, less candid, or otherwise less engaged with the material than White faculty members who more closely track White American cultural norms.¹⁰⁰

Hand gestures and body movement also differ significantly by race and culture. African Americans, on average, use more intense body language than White Americans do.¹⁰¹ Conversely, Chinese Americans, Japanese Americans, and Korean Americans use less expressive body language than Whites, and display their emotions less visibly.¹⁰² Japanese Americans are also less assertive than White Americans during verbal interactions.¹⁰³ Disparities like these can prompt White students, still the majority in most law school classrooms, to view African American

⁹⁸ Uw Gielen, et al., *Naturalistic Observation of Sex and Race Differences in Visual Interactions*, 9 Int'l J. Group Tensions 211 (1979); Marianne LaFrance & Clara Mayo, *Racial Differences in Gaze Behavior During Conversations: Two Systematic Observational Studies*, 33 J. Personality & Soc. Psychol. 547 (1976). See also Kyung Soon Lee & Angela Carrasquillo, *Korean College Students in the United States: Perceptions of Professors and Students*, 40 College Student J. 442 (2006) (White professors perceived that Korean students avoided eye contact during conversations).

⁹⁹ Anjanie McCarthy et al., *Cultural Display Rules Drive Eye Gaze During Thinking*, 37 J. Cross-Cultural Psychol. 717 (2006).

¹⁰⁰ Conversely, as the percentage of nonwhite and international students grows among law students, White faculty may find themselves rated negatively by students offended by their eye contact. Cf. Babad, *supra* note 97, at 290 (“in some cultures, looking someone straight in the eye is not considered positive at all, but rather aggressive, daring, and impolite”); Judith A. Sanders & Richard L. Wiseman, *The Effects of Verbal and Nonverbal Teacher Immediacy on Perceived Cognitive, Affective, and Behavioral Learning in the Multicultural Classroom*, 39 Commun. Educ. 342, 351 (2003) (for African American students, instructor eye contact was not related to perceived cognitive learning; eye contact was significant for Asian, White, and Hispanic students).

¹⁰¹ Thomas Kochman, *Force Fields in Black and White Communication*, in *Cultural Communication and Intercultural Contact* 193 (Donal Carbaugh ed. 1990). See also Sanders & Wiseman, *supra* note 100, at 351 (an instructor’s tense body position correlated with White students’ perceived learning, but not with that of Black, Asian, or Hispanic students); Stella Ting-Toomey, *Conflict Communication Styles in Black and White Subjective Cultures*, in *Interethnic Communication: Current Research* 75 (Young Yun Kim ed. 1986) (Blacks showed more confrontational conflict styles than Whites did).

¹⁰² J.M. Jones, *Racism in Black and White: A Bicultural Model of Reaction and Evolution*, in *Eliminating Racism: Profiles in Controversy* 118 (P.S. Katz & D.A. Taylor eds. 1988); Min-Sun Kim, *A Comparative Analysis of Nonverbal Expressions as Portrayed by Korean and American Print-Media Advertising*, in *Readings in Cultural Contexts* 206 (Judith N. Martin et al. eds. 1998).

¹⁰³ William B. Gudykunst et al., *Uncertainty Reduction in Japanese-American/Caucasian Relationships in Hawaii*, 51 W.J. Speech Commun. 256, 269 (1987).

professors as more hostile than White ones, while they view Asian American professors as cold, uncaring, or diffident.

Recent research, finally, suggests that sexual orientation also shapes nonverbal behavior. Using the “thin slice” method described above,¹⁰⁴ Nalini Ambady and two colleagues found that undergraduates correctly identified another person’s sexual orientation about 70% of the time after viewing just ten seconds of silent videotape.¹⁰⁵ Students may make similar inferences about a faculty member’s sexual orientation based on nonverbal cues; those inferences can affect evaluations in a variety of ways. If heterosexual students are uncomfortable with gays and lesbians, and they correctly identify a professor’s homosexual orientation through his or her nonverbal behavior, that professor may receive more negative evaluations. Alternatively, since judgments of sexual orientation are imperfect, some students may erroneously attribute a particular orientation to a professor and respond negatively to the gestures signifying that orientation. Research on the intersection of sexual orientation, nonverbal behaviors, and social judgments is just beginning, but may uncover significant concerns about teaching evaluations.

These traits constitute only some of the many nonverbal behaviors that influence judgments that students make about faculty.¹⁰⁶ The pervasive influence of these behaviors, combined with their immutability, explains the difficulty that many dedicated professors have encountered in trying to raise their scores on student evaluations. Although training has improved ratings for some faculty,¹⁰⁷ most professors realize modest gains at best. Even twenty weeks of

¹⁰⁴ See *supra* notes 43-57 and accompanying text.

¹⁰⁵ Nalini Ambady et al., *Accuracy of Judgments of Sexual Orientation from Thin Slices of Behavior*, 77 *J. Personality & Soc. Psychol.* 538, 541 (1999).

¹⁰⁶ See also Hillary Anger Elfenbein & Nalini Ambady, *On the Universality and Cultural Specificity of Emotion Recognition: A Meta-Analysis*, 128 *Psychol. Bull.* 203 (2002) (reviewing studies of cross-cultural emotion recognition); Donald L. Rubin, *Help! My Professor (Or Doctor or Boss) Doesn't Talk English!*, in *Readings in Cultural Contexts* 149 (Judith N. Martin et al. eds. 1998) (discussing weak tolerance most American students have for any variations in accent, even within their own culture).

¹⁰⁷ See, e.g., Williams & Ceci, *supra* note 6.

professional instruction attended by one group of highly motivated faculty generated just a small increase in student evaluation scores.¹⁰⁸ Anecdotal reports express similar frustrations; professors, for example, are puzzled to discover that expanding office hours and giving students detailed contact information does not improve their ratings on “accessibility.”¹⁰⁹ As two reviewers concluded, “accessibility” as measured on student evaluations “has more to do with personality than office hours.”¹¹⁰ And personality, especially as reflected through unconscious mannerisms, is notoriously hard to change.

B. *Perceptual Filters: Stereotyping*

A faculty member’s gestures, voice tones, facial expressions, and other nonverbal behaviors profoundly shape what students believe about that professor’s teaching effectiveness. The professor’s actual behaviors, however, tell only half the story. Law students, like all humans, perceive other people’s behavior through filters that are socially conditioned; none of us see the world through neutral, objective lenses. Instead, our minds classify individuals according to race, gender, age, and other socially salient categories with dizzying speed.¹¹¹ We then use those classifications to interpret a speaker’s behavior, so that the same gestures, expressions, and other

¹⁰⁸ Harry Murray & C. Lawrence, *Speech and Drama Training for Lecturers as a Means of Improving University Teaching*, 13 *Research in Higher Educ.* 73 (1980); Murray, *Classroom Teaching*, *supra* note 32, at 31. A professional actress taught the faculty members breathing and voice exercises, directed them in enacting short dramatic scenes, and gave them corrective feedback on their lectures; the group met for two hours each week. The faculty members who participated in the training averaged a 0.2 gain in student ratings (on a 5-point scale) while a control group of faculty, who did not take part in the training sessions, realized no gains. An earlier study by one of Murray’s graduate students identified even less pay-off from more modest attempts to provide feedback to professors on their classroom behaviors, although the lowest rated instructors realized some gains from that method. *Id.* at 29-31.

¹⁰⁹ Clayson & Haley, *supra* note 42, at 13.

¹¹⁰ *Id.*

¹¹¹ Ambady et al., *supra* note 43, at 231 (citing studies); Tiffany A. Ito et al., *The Social Neuroscience of Stereotyping and Prejudice: Using Event-Related Potentials to Study Social Perception*, in *Social Neuroscience: People Thinking About People* 189 (John T. Cacioppo et al. eds. 2006). Kathleen Bean describes the power of these classifications in the context of a law professor: “[T]he gender gap . . . is born the moment I walk into the classroom. It has a life of its own before I open my mouth, before my body language speaks, and before my eyes make contact with anyone. My sex, and my sex alone, . . . opens the gender gap.” Bean, *supra* note 5, at 29.

components of nonverbal behavior look different depending on the speaker's race, gender, and other characteristics.

A long line of research, for example, demonstrates that Americans perceive smiles and friendliness more readily on White faces than on African American ones; conversely, Americans more readily perceive anger and hostility in Black individuals. Birt Duncan conducted one of the earliest studies in this area, showing a series of college students one of four videotapes in which a male student shoved another student after a heated disagreement.¹¹² The tapes systematically varied the races of the disputing students, producing dramatically different responses among viewers. When a Black student shoved a White peer, 75% of the viewers described the incident as "violent." When a White student shoved a Black, only 17% of the viewers registered the same reaction. Instead, the students characterized this behavior as "playing around," "dramatizing," or "aggressive."¹¹³ The race of the perpetrator, race of the victim, and interaction between the two significantly affected the viewers' perceptions.¹¹⁴

More recently, a research team led by Joshua Correll used a laboratory computer game to demonstrate dramatically different reactions to the identical behavior displayed by White and

¹¹² Birt L. Duncan, *Differential Social Perception and Attribution of Intergroup Violence: Testing the Lower Limits of Stereotyping Blacks*, 34 J. Personality & Soc. Psychol. 590 (1976). The subjects believed they were viewing a live encounter between two students in a nearby room, using a closed circuit television. To standardize interactions, however, the interactions were taped.

¹¹³ *Id.* at 595. Subjects chose these descriptions from standard rating scales. Their evaluation of this behavior was embedded in a larger exercise in which the subjects rated several incidents.

¹¹⁴ The White viewers characterized a Black student shoving another Black as the most violent; a Black shoving a White as next most violent; White shoving a Black next; and a White shoving a White as most benign. *Id.* at 597. Race also affected the attributions that viewers attached to the students' behavior. Viewers attributed shoving by the Black student to the student's personal attributes (e.g., his violent nature), while they attributed shoving by a White to circumstantial factors (i.e., a stimulus caused the student to act). *Id.* at 597.

Sagar and Schofield obtained similar results in a study of sixth grade boys. H. Andrew Sagar & Janet W. Schofield, *Racial and Behavioral Cues in Black and White Children's Perceptions of Ambiguously Aggressive Acts*, 39 J. Personality & Soc. Psychol. 590 (1980). When asked to judge ambiguous behavior in a hypothetical story, the boys rated the actions of Black children as significantly more mean or threatening than the identical actions by White children. *Id.* at 594.

Black actors.¹¹⁵ Correll and his colleagues constructed a game in which a series of men appeared on the screen; the researchers systematically varied the race of the men, the poses they adopted, and the objects they held. Game players were instructed to “shoot” men holding guns and “not shoot” those holding innocent objects like wallets.¹¹⁶ The players proved significantly more likely to shoot Blacks holding innocent objects than Whites brandishing the same items.¹¹⁷ They also showed a greater tendency to overlook (“not shoot”) Whites holding guns than Blacks wielding them.¹¹⁸ And even when subjects accurately distinguished among targets, they took significantly longer to recognize Whites with guns and Blacks with innocent objects than vice versa.¹¹⁹

These stereotypes affect the judgments of both Black and White subjects; Correll and several other researchers have shown that both Whites and Blacks perceive Black men as more dangerous than White men engaged in identical actions.¹²⁰ Americans of both races unconsciously filter their perceptions of nonverbal behavior to see more aggression among Blacks than among Whites engaged in the same conduct.

Several scholars, moreover, have shown that these differences arise even in very subtle contexts. In a striking series of experiments, Kurt Hugenberg demonstrated that White students identify happy expressions on White faces significantly more quickly than they see sad or angry

¹¹⁵ Joshua Correll et al., *The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals*, 83 J. Personality & Soc. Psychol. 1314 (2002).

¹¹⁶ One computer key corresponded to “shoot,” while another indicated “don’t shoot,” so subjects had to register a reaction one way or another. *Id.* at 1316.

¹¹⁷ *Id.* at 1318 (Study One), 1319 (Study Two), 1322 (Study Three).

¹¹⁸ *Id.* at 1319 (Study Two). These results did not reach significance in the other studies, although they showed the same direction.

¹¹⁹ *Id.* at 1317 (Study One), 1322 (Study Three). The window for response times in Study Two was so small that differences in response latencies were not detected. *See also* Joshua Correll et al., *Event-Related Potentials and the Decision to Shoot: The Role of Threat Perception and Cognitive Control*, 42 J. Experimental Psychol. 120 (2006) (replicating results with yet another group of students); Justin D. Levinson, *Forgotten Racial Equality: Implicit Bias, Decision-Making, and Misremembering* (Aug. 25, 2006), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=927547 (subjects were more likely to remember aggressive acts attributed to African American or Hawaiian males than to White ones).

¹²⁰ Correll et al., *supra* note 115, at 1324 (Study Four); Sagar & Schofield, *supra* note 114, at 594-95.

expressions on those faces.¹²¹ The opposite pattern emerges when White students view Black faces; they identify sad or angry faces on Blacks significantly more rapidly than they see happy expressions on those faces.¹²²

Even neutral expressions elicit different responses depending on the race of the target. Pierre Philippot and Yanéla Yabar showed White students a series of photographs that had been carefully selected for their neutral expressions.¹²³ The students were significantly more likely to associate neutral Black faces, rather than neutral White ones, with emotions like “show[s] aggressiveness to others,” “insult[s] others,” and “boil[s] inwardly.”¹²⁴

These studies confirm that “how individuals perceive and categorize facial expressions can depend quite critically on who it is that is displaying the expression.”¹²⁵ In the classroom, students will detect warm, happy faces more quickly on White professors while they see sad, angry, or hostile expressions more readily on Blacks. A neutral expression on the face of a White professor

¹²¹ Kurt Hugenberg, *Social Categorization and the Perception of Facial Affect: Target Race Moderates the Response Latency for Happy Faces*, 5 *Emotion* 267 (2005). Students viewed the faces on computer monitors, using keystrokes to identify the expressed emotion. Hugenberg used character animation software to construct male faces that shared identical facial characteristics, varying only in skin color and expression. Pretesting confirmed that students readily identified the expressions as happy, sad, or angry. In the main studies, each student responded to 160 trials using eight stimulus faces in random order. *Id.* at 271-72.

¹²² *Id.* Hugenberg and a colleague determined that the effect is even greater among White students who show higher degrees of racial prejudice on the Implicit Attitude Test. Kurt Hugenberg & Galen V. Bodenhausen, *Facing Prejudice: Implicit Prejudice and the Perception of Facial Threat*, 14 *Psychol. Sci.* 640 (2003). The difference, however, occurs among even less prejudiced students. See also Kurt Hugenberg & Galen V. Bodenhausen, *Ambiguity in Social Categorization: The Role of Prejudice and Facial Affect in Race Categorization*, 15 *Psychol. Sci.* 342 (2004).

¹²³ Pierre Philippot & Yanéla Yabar, *Stereotyping and Action Tendencies Attribution as a Function of Available Emotional Information*, 35 *European J. Soc. Psychol.* 517 (2005). Philippot and Yabar conducted their study in Belgium, which has a significant North African immigrant population that has given rise to racial tensions and stereotypes similar to those between Whites and Blacks in the United States. In particular, White Belgians “generally perceive[]” North African immigrants “as a threatening and aggressive group.” *Id.* at 520.

¹²⁴ *Id.* at 524 (Study One), 527 (Study Two). Philippot and Yabar tested two other stereotypes of North Africans: “Show exuberance” and “Show excitement.” *Id.* at 522. It is less clear that these stereotypes mark Black/White relationships in the United States. Unfortunately, Philippot and Yabar did not distinguish among reactions to these five different stereotypes.

The laboratory studies examining differential perceptions of hostility on Black and White faces have been limited so far to male faces and actors. Hopefully researchers will expand these investigations to explore perceptions of Black and White women. Meanwhile, accounts by Black women faculty suggest that they, like Black men, suffer from exaggerated perceptions of hostility. See, e.g., Smith, *supra* note 1, at 114 (discussing the stereotype of Black women as “Sapphire,” a “tough, domineering, emasculating, strident and shrill” character).

¹²⁵ Hugenberg, *supra* note 121, at 275.

may convey warmth to students, while the same expression on a Black professor's face connotes hostility.¹²⁶ These differences can markedly affect student evaluations of faculty accessibility, caring, and other qualities.¹²⁷

Perceptual filters also bias the ways in which people perceive identical behaviors of men and women. When a male employee establishes direct eye contact with another worker, he raises his credibility.¹²⁸ A female employee using the same behavior does not enhance her credibility; instead, she increases the perception that she will act coercively against the other worker.¹²⁹ Similarly, a relaxed facial expression in male employees connotes credibility, expert knowledge, legitimate power, and the ability to confer rewards on others.¹³⁰ A relaxed expression on the face

¹²⁶ Anecdotal evidence from campuses yields similar conclusions. A Black student on one predominantly White campus noted: "Socially, Blacks are pressed into being very passive and always grinning, otherwise they are immediately typecast as being hostile and aggressive." Carole Baroody Corcoran & Aisha Renée Thompson, "What's Race Got to Do, Got to Do with It?" *Denial of Racism on Predominantly White College Campuses*, in 1 *Racism in America: The Psychology of Prejudice and Discrimination* 142 (Jean L. Chin ed. 2004).

¹²⁷ Much of the research on race biases, like this example, focuses on differences between Blacks and Whites. Analogous differences, however, mark comparisons of other race groups; research is starting to illuminate these complex relations. For recent studies of other racial groups, see, e.g., Glascock & Ruggiero, *supra* note 3 (analyzing evaluations of White and Hispanic professors by students of both ethnicities); Levinson, *supra* note 119 (examining recall bias as applied to African Americans, Whites, and Hawaiians); Dominic W. Massaro & John W. Ellison, *Perceptual Recognition of Facial Affect: Cross-Cultural Comparisons*, 24 *Memory & Cognition* 812 (1996) (comparing White American and native Japanese students).

¹²⁸ Herman Aguinis et al, *Effects of Nonverbal Behavior on Perceptions of Power Bases*, 138 *J. Soc. Psychol.* 455 (1998). Aguinis and his colleagues asked 170 young adults with significant workplace experience to respond to a written vignette in which two employees, "John" and "Greg," discussed declining profits at their bank. The researchers systematically varied John's eye contact with Greg, his facial expression, and his body posture in the descriptions to capture a variety of nonverbal behaviors. *Id.* at 460. When the vignette described John as looking directly at Greg, readers rated John's credibility as significantly higher than when the vignette indicated that John looked around the room, glancing occasionally at Greg. *Id.* at 463. Credibility was assessed from reactions to a series of statements such as "John is a man who keeps his word" and "John tells the truth." *Id.* at 462.

¹²⁹ Herman Aguinis & Christine A. Henle, *Effects of Nonverbal Behavior on Perceptions of a Female Employee's Power Bases*, 141 *J. Soc. Psychol.* 537, 544-46 (2001). Aguinis and Henle used the same procedure adopted in Aguinis' earlier study, see *supra* note 128, but identified the two employees as "Mary" and "John." When Mary was described as looking directly at John, readers did not believe that she was more credible than when she glanced at him only occasionally. Instead, they perceived her as possessing significantly more coercive power. *Id.* at 543-44. Coercive power indicates a supervisor's ability to punish a subordinate; Aguinis and Henle measured it through reactions to statements like "Mary can give John undesirable job assignments" and "Mary can make things unpleasant on the job for John." *Id.* at 543.

¹³⁰ Aguinis et al., *supra* note 128, at 463. This research, more fully described *supra* note 128, asked young adults to assess vignettes of two male employees interacting in the workplace. When the vignette described the facial expression of one worker, "John," as relaxed, readers rated him significantly higher on credibility, expert power, legitimate power, and reward power—all positive attributes in the workplace. The assessment of credibility is described *supra* note 128. Expert power measures the special expertise that a superior employee can share with an inferior; Aguinis and his colleagues measured it through reactions to statements like "John can share with Greg his

of a male manager also suggests that the manager can make subordinates feel valued, approved, and important.¹³¹ Calm expressions on the faces of female employees confer none of these positive benefits. Instead, a relaxed expression prompts observers to decrease their estimation of the woman's power.¹³²

These biases, like those based on racial categories, can powerfully affect students' evaluations of male and female professors. Male instructors who establish eye contact with students and adopt a relaxed expression—attitudes typically effective in classroom teaching—appear credible, knowledgeable, and in control of the classroom. The students will believe that the professor has legitimate authority to demand work from them and will focus on his power to reward them for good answers, rather than punish them for bad ones. They will also see the professor's ability to make them feel valued, approved, and important.

Female faculty who adopt the same classroom demeanors will evoke far different responses. They are less likely to gain credibility or the appearance of expert knowledge in students' eyes. Students will be less likely to acknowledge the female faculty member's authority to make assignments or demand work in the same way they will perceive that power in a male professor. They will also be less likely to focus on the female professor's ability to reward them

(John's) considerable experience and/or training." Aguinis et al., *supra* note 128, at 462. Legitimate power refers to an employee's perceived authority to command others. It was measured in this study through reactions to statements like "John can make Greg realize that he (Greg) has tasks to accomplish." *Id.* Reward power, finally, consists of the ability to confer benefits on another worker. This study established perceptions of reward power through reactions to statements like "John can increase Greg's pay level." *Id.*

¹³¹ *Id.* at 463. In this study, described further *supra* notes 128-30, a relaxed expression on a male manager's face significantly increased ratings of "referent power." That type of power includes the perceived ability to make others feel "valued," "approve[d]," "personally accepted," and "important." Aguinis et al., *supra* note 128, at 462.

¹³² Aguinis & Henle, *supra* note 129, at 544-46. Aguinis and Henle identified just one nonverbal behavior that appeared to have a positive effect on perceptions of female managers. A relaxed body position, signified by leaning back in a chair with legs crossed, increased ratings of referent power. *Id.* at 545. Readers, in other words, were more likely to think that this female employee could confer feelings of acceptance, value, and importance on another worker. Aguinis and Henle termed this finding "surprising" and "did not consider this a finding of strong theoretical significance" because of its incongruity with prior theoretical work. *Id.* at 545. The finding, however, may hold a clue to how female professors *do* achieve positive ratings from students. When they appear relaxed in the classroom, that appearance may connote high status (consistent with a general correspondence between those variables), which for women translates into an ability to confer "soft" benefits like feelings of value, acceptance, and importance.

for good work or to enhance their feelings of value and importance. Instead, students are more likely to focus on the female instructor's power to punish them for poor work.¹³³

Fewer studies examine the impact of socioeconomic status on perceptions, but those that exist suggest that this category also shapes attitudes. One well developed line of experiments uses vehicles—a common class marker in American life—to probe reactions to people of high and low socioeconomic class. Anthony Doob and Alan Gross performed the first of these experiments, demonstrating that drivers were significantly more likely to honk at an old, rusty car that failed to move promptly through an intersection than at an expensive, new, and well maintained car.¹³⁴ Andrew McGarva and Michelle Steiner observed similar differences in driver responses to a rusty Ford pickup (low-status vehicle) and new Nissan Pathfinder SUV (high-status vehicle).¹³⁵ Drivers

¹³³ This type of power is a negative one leading to “resistance . . . , lower organizational commitment, and dissatisfaction.” *Id.* at 544-45. Several studies comparing student comments on teaching evaluations reveal gender differences parallel to these differences. *See, e.g.,* Farley, *supra* note 5, at 339 (1996) (evaluations of female law faculty criticize them disproportionately for “being too strict or for being ‘task-masters’”); Michael A. Messner, *White Guy Habitus in the Classroom, 2 Men & Masculinities* 457, 458 (2000) (evaluations noted that male co-teacher was “objective,” “looks at all sides of an issue,” “relaxed and comfortable,” “flexible,” “open-minded,” and “good humored,” while a female colleague was perceived as “biased,” having “an agenda,” having a “chip on her shoulder,” “rigid and dogmatic,” “politically correct,” “grumpy and angry”).

Empirical surveys suggest that female faculty do, on average, receive lower evaluations than their male colleagues. A recent, multivariate analysis of almost 17,000 student evaluations at the University of Texas revealed significantly higher evaluations for male faculty, even after controlling for course type and instructor status. Hamermesh & Parker, *supra* note 3. Another recent study of evaluations gathered in 741 different courses taught at 21 different institutions showed that women faculty received significantly lower ratings from male students than female ones; conversely, the male and female students did not differ in their assessment of male faculty. John A. Centra & Noreen B. Gaubatz, *Is There Gender Bias in Student Evaluations of Teaching?*, 71 *J. Higher Educ.* 17 (2000). Earlier studies reached mixed results, but these most recent, and most comprehensive, studies suggest that male and female faculty receive somewhat different ratings.

¹³⁴ Anthony N. Doob & Alan E. Gross, *Status of Frustrator as an Inhibitor of Horn-Honking Responses*, 76 *J. Soc. Psychol.* 213 (1968). The drivers also honked more quickly at the low status cars and were more likely to honk repeatedly at those cars. *Id.* at 215-16. Andreas Diekmann and several colleagues complemented this research by finding that the status of the *blocked* car also affects responses in these situations. Drivers of higher status cars are more likely to honk their horns or flick their headlights at the waiting car than are drivers of lower status cars. Andreas Diekmann et al., *Social Status and Aggression: A Field Study Analyzed by Survival Analysis*, 136 *J. Soc. Psychol.* 761 (1996).

¹³⁵ Andrew R. McGarva & Michelle Steiner, *Provoked Driver Aggression and Status: A Field Study*, *Transportation Research Part F* 3, at 167.

accelerated away from the Ford pickup significantly more quickly than they drove away from the SUV after the other driver honked at them.¹³⁶

Law students don't honk at professors who displease them, but the same attitudinal differences can affect relationships in the classroom. The "horn honking" studies expose a cultural tendency to vent frustration or hostility more readily against low-status individuals than high-status ones. Socratic classrooms, challenging material, and intense competition for grades are at least as frustrating to students as a stalled car at an intersection. Law students may express that irritation more readily on evaluations of professors with low-status mannerisms than in their assessments of faculty with more high-status appearances. Indeed, the horn-honking studies may explain the surprising degree of overt hostility that law students express on evaluations of some minority faculty.¹³⁷ Those evaluations are a type of classroom "honking."

These examples comprise just a sample of the kind of stereotypes that routinely bias perceptions based on race, gender, class, and other categories. Researchers have documented a substantial number of other biases, with still others to be uncovered.¹³⁸ The bottom line is that, especially when the mind makes judgments based on "thin slices" of nonverbal behavior, as students apparently do in teaching evaluations, these stereotypes automatically and unconsciously

¹³⁶ *Id.* at 173-74. The "provoking" driver also held "his mouth expressively agape" and "raised both arms impatiently." *Id.* at 172. See also Kay K. Deaux, *Honking at the Intersection: A Replication and Extension*, 84 J. Soc. Psychol 159 (1971) (finding a status difference in the same direction as that identified by Doob and Gross, but that did not reach statistical significance; status differential of cars was not as great). There is also evidence that drivers respond more aggressively towards female drivers who frustrate them than male ones. *Id.* Status influences on the road, however, may differ more among cultures than some other behaviors do. See, e.g., Diekmann et al., *supra* note 134, at 768

¹³⁷ See *supra* note 4 and accompanying text. See also Smith, *supra* note 1, at 179 (quoting anonymous note received by an African American female professor from a law student) ("A black mammy like you is completely incompetent to judge anyone on anything. I do not care whether you are magna cum laude. . . . You do not belong in law school teaching. Black mammies should stay at home doing mammy things. Or they should stay in their place and it is not law school.").

¹³⁸ See, e.g., Ron Tamborini & Dolf Zillman, *College Students' Perception of Lecturers Using Humor*, 52 Perceptual & Motor Skills 427 (1981) (students respond differently to male and female professors using similar humor); Marc David Pell, *Evaluation of Nonverbal Emotion in Face and Voice: Some Preliminary Findings on a New Battery of Tests*, 12 Tennessean 499 (2002) (subjects recognized neutral expressions more readily on male faces and expressions of disgust more readily on female ones).

alter perceptions. Students, like other humans, ascribe different meaning to the same behaviors depending on the race, gender, class, and other characteristics of the actor.

Acquaintance does not necessarily reduce these perceptual distortions; instead, the differences may grow through self-reinforcing cycles. If students perceive hostility in the ambiguous expression of a Black professor, coerciveness in the eye contact of a woman, or obstructive behavior in a professor from an underprivileged background, they will respond to those impressions with heightened hostility of their own. The faculty member, in turn, confronts negative reactions that have no apparent source; he or she has engaged in neutral (or even positive) behaviors that appear to have provoked an angry response. Faced with this seemingly irrational hostility, the Black, female, or low-status faculty member may display subtle signs of their own discomfort and anxiety. Students will perceive those behaviors, reinforcing their initial, negative images. False impressions generated by stereotypes, in other words, create a cycle of mutually reinforcing behavior.¹³⁹ By the semester's end, students may feel quite justified in rating the Black, female, or low-status professor as hostile and uncaring. Because humans are so sensitive to nonverbal signals of aggression, support, and other attitudes, even small differences in those perceptions have "powerful implications" for ongoing interactions.¹⁴⁰

C. *The Missing Link with Learning*

The link between nonverbal behavior and teaching evaluations allows irrelevant characteristics like race and gender to distort student assessments of faculty. The same association hints that conventional evaluations bear little relationship to student learning. The professionals

¹³⁹ See Hugenberg & Bodenhausen, *supra* note 122, at 643; C.O. Word et al., *The Nonverbal Mediation of Self-Fulfilling Prophecies in Interracial Interaction*, 10 J. Experimental Soc. Psychol. 109 (1974); see also Chamallas, *supra* note 5, at 202 (describing such a cycle when students respond critically to a female law professor); Lu-in Wang, *Race as Proxy: Situational Racism and Self-Fulfilling Stereotypes*, 53 DePaul L. Rev. 1013 (2004) (discussing the phenomenon of self-fulfilling stereotypes in a variety of legal contexts).

¹⁴⁰Hugenberg & Bodenhausen, *supra* note 122, at 643.

who lauded Dr. Fox's entertaining but meaningless discussion of mathematical theory and medical education could not have learned much: there was nothing in the lecture to learn.¹⁴¹ Behaviors detectable in just thirty seconds of silent videotape seem unlikely to promote solid learning, yet these elements strongly influence student evaluations.¹⁴² Do student evaluations of faculty correlate with student learning?

The cumulative research suggests that there is little, if any, positive association between the ratings students give faculty and the amount they learn. The most recent study, in fact, suggests a negative correlation between evaluations and learning. In a particularly well designed investigation, two business professors gathered eight full years of data on students who completed two sequential accounting courses at a Midwestern university.¹⁴³ After controlling for ACT scores, overall GPA, and grade in the first course, the researchers discovered that students who completed the first course with highly rated professors achieved significantly *lower* grades in the second course.¹⁴⁴ The professors with top evaluations, in other words, did not prepare students for the more advanced course as well as lower rated faculty did.

This study is particularly noteworthy because it captured multiple facets of learning. By examining performance in a subsequent course, one that built directly on material learned in the first class, the investigation tapped students' long-term retention of material, as well as their

¹⁴¹ See *supra* notes 22-27 and accompanying text.

¹⁴² See *supra* notes 45-57 and accompanying text. See also Clayson & Sheffet, *supra* note 6, at 157-58 ("Finding an association between the final evaluation of the class at the end of the term and a personality evaluation made within 5 minutes of exposure . . . makes a validity argument for the relationship between personality and evaluation difficult to defend.").

¹⁴³ Penelope Yunker. & James Yunker, *Are Student Evaluations of Teaching Valid? Evidence from an Analytical Business Core Course*. 78 J. Educ. for Business 313 (2003).

¹⁴⁴ *Id.* at 316. ACT score (which provided a measure of pre-college aptitude), overall GPA, and introductory course grade, on the other hand, all showed significant positive correlations with the student's grade in the intermediate course. As one would expect, the students' general academic ability (as reflected by ACT score and overall GPA), their diligence (as reflected by overall GPA), and their relative performance in the initial course all predicted achievement in the subsequent course. The notable finding of the Yunkers' study is that, after controlling for these variables, students who learned the initial material from more highly rated professors performed less well than students who studied that material from professors with lower evaluations.

motivation for learning. Some scholars suggest that, even if highly rated professors do not enhance their students' immediate achievement, they inspire an enthusiasm for learning that has longer term pay-off.¹⁴⁵ In the Yunkers' study, however, this relationship did not hold. Students who completed the initial course with highly rated professors performed less well than students who learned the introductory material from lower rated faculty. The results of this analysis lend support to speculation that a desire for better teaching evaluations may pressure faculty to “dumb down” the material they present to students.¹⁴⁶

In several other studies, teaching evaluations failed to show any correlation—positive or negative—with student achievement. An examination of students in two sequential economics courses failed to find any significant relationship between faculty evaluations in the first course and student performance in the advanced offering.¹⁴⁷ Professors rated poorly by students nonetheless prepared them as effectively as more highly rated colleagues. Another, more recent study directly probed the relationship between the nonverbal behaviors that influence teaching evaluations and student performance in a course. Students enrolled in fifteen different sections of an introductory course scored their instructor's nonverbal behaviors shortly before completing a common midterm exam.¹⁴⁸ Scores on the midterm showed no significant correlation with the students' ratings of their instructors.¹⁴⁹

¹⁴⁵ See, e.g., Mike Allen et al., *Learning: Using Meta-Analysis to Test a Causal Model*, 55 *Commun. Educ.* 21 (2006); Paul L. Witt et al., *A Meta-Analytical Review of the Relationship Between Teacher Immediacy and Student Learning*, 71 *Commun. Monographs* 184 (2004).

¹⁴⁶ See, e.g., Johnson, *supra* note 6.

¹⁴⁷ Stephen Shmanske, *On the Measurement of Teacher Effectiveness*, *Research in Econ. Educ.*, Fall 1988, at 307. See also Ganesh Mohanty et al., *Multi-Method Evaluation of Instruction in Engineering Classes*, 18 *J. Personnel Evaluation Educ.* 139 (2005) (finding no significant relationship between student learning, as measured by pre- and post-tests, and student ratings of instruction in engineering courses, although study included a small number of students).

¹⁴⁸ Jon A. Hess et al., *Is Teacher Immediacy Actually Related to Student Cognitive Learning?*, 52 *Communication Studies* 197, 199 (2001).

¹⁴⁹ *Id.* at 210. See also Joseph L. Chesebro, *Effects of Teacher Clarity and Nonverbal Immediacy on Student Learning, Receiver Apprehension, and Affect*, 52 *Commun. Educ.* 135 (2003); Debra Q. O'Connell & Donald J. Dickinson, *Student Ratings of Instruction as a Function of Testing Conditions and Perceptions of Amount Learned*, 27

Stephen Ceci, the psychology professor who documented his success in substantially raising student evaluations by adopting minor changes in his nonverbal behavior, offers further evidence that student evaluations bear little relationship to learning. After implementing a media consultant's advice, Ceci's evaluations rose almost a full point on a 5-point scale.¹⁵⁰ The students enrolled in the post-training course, moreover, believed that they learned more from a professor who gestured emphatically and varied his voice tones. On a 5-point scale, these students awarded Ceci an average of 4.05 for how much they had learned, compared to just 2.93 registered by students the previous semester.¹⁵¹ The media training and changes in nonverbal behavior, however, did not affect students' actual performance on quizzes and exams; those scores were virtually identical across the two semesters.¹⁵² Ceci's changed style, in other words, improved his evaluations but did not enhance the students' learning.

A few studies do identify a positive relationship between student evaluations and learning. Harry Murray, for example, found a correlation of .30 between students' ratings of their instructor and final exam scores in a study of multiple sections of an introductory psychology course.¹⁵³ Differences in evaluation scores, in other words, explained about 9% of the variance in student learning.¹⁵⁴ An earlier meta-analysis by Peter Cohen identified a somewhat higher correlation of .43 between teaching evaluations and student learning, suggesting that the former explain about

J. Research & Dev. Educ. 18 (1993) (student ratings of instructor in education course failed to correlate with amount learned as measured by pre- and post-tests).

¹⁵⁰ See *supra* notes 34-37 and accompanying text.

¹⁵¹ Williams & Ceci, *supra* note 6. This finding is consistent with a large number of studies finding that student evaluations of teaching correlate highly with students' perceptions of how much they have learned. See, e.g., O'Connell & Dickinson, *supra* note 149; Witt et al., *supra* note 145 (summarizing studies). Students, in other words, believe that they learn more from highly rated professors—although they usually do not.

¹⁵² Williams & Ceci, *supra* note 6.

¹⁵³ *Effective Teaching Behaviors*, *supra* note 32, at 151. See also Jeff Koon & Henry G. Murray, *Using Multiple Outcomes to Validate Student Ratings of Overall Teacher Effectiveness*, 66 J. Higher Educ. 61 (1995) (analyses of same data).

¹⁵⁴ For an explanation of the concepts of correlation and variance, see *supra* note 51.

18.5% of the variance in the latter.¹⁵⁵ Researchers, however, have noted several flaws in Cohen's analysis: It included studies in which students rated their professors only after receiving course grades, as well as surveys based solely on students' subjective beliefs about the extent of their learning.¹⁵⁶ In light of these and other problems, Cohen joined two other evaluation experts in agreeing that additional research was needed.¹⁵⁷ Those more recent studies, as noted above, have shown no positive relationship between teaching evaluations and student learning.

In a meta-analysis of eighty-one studies published between 1979 and 2001, finally, Paul Witt and several colleagues identified a small positive relationship between "immediacy," the complex of personality and nonverbal behaviors that produce high student evaluations, and objective measures of student achievement like exam scores.¹⁵⁸ Given the relationship between immediacy and evaluations, this analysis could signal a parallel association between evaluations and learning. The correlation calculated by Witt and his colleagues, however, was just .122,¹⁵⁹ meaning that variation in the measured nonverbal behaviors explained just 1.5% of the variance in student learning. Although statistically significant, Witt and his colleagues concluded that these "low associations" were "less than meaningful" in any practical sense.¹⁶⁰

¹⁵⁵ Peter A. Cohen, *Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies* 51 Rev. of Educ. Research 281 (1981).

¹⁵⁶ See, e.g., Philip C. Abrami et al., *Validity of Student Ratings of Instruction: What We Know and What We Do Not*, 82 J. Educ. Psych. 219 (1990); Hinton, *supra* note 7. The former problem confounds attempts to correlate student achievement with faculty ratings, because a high evaluation offered after receiving course grades may reflect student satisfaction with their grades rather than an independent judgment of faculty quality. See also *supra* note 9 (discussing the relationship between student grades and teaching evaluations). The second flaw substantially diluted Cohen's conclusions; most scholars consider student perceptions of learning a poor measure of their actual learning. See, e.g., Witt et al., *supra* note 145, at 189-90.

¹⁵⁷ Abrami et al., *supra* note 156, at 231.

¹⁵⁸ Witt et al., *supra* note 145, at 185.

¹⁵⁹ *Id.* at 196. This figure represents the correlation of student learning with combined measures of verbal and nonverbal immediacy. Attempts to separate verbal and nonverbal immediacy yielded even lower correlations. "Cognitive learning" embraces all types of learning that law faculty attempt to convey. *I.e.*, it includes thinking and analysis skills, as well as mastery of basic facts and principles.

¹⁶⁰ *Id.* at 160. Social science conventions agree that correlations of this magnitude are weak, with little practical significance. See *supra* note 51.

The nonverbal behaviors that students reward on teaching evaluations, in other words, produce little if any pay off in student learning. Those behaviors seem to make learning more enjoyable for students, and that is a worthwhile goal for faculty to pursue. To the extent that nonverbal behaviors are mutable, professors can attempt to learn mannerisms that will increase students' enjoyment of their classes as well as the faculty member's teaching evaluations. Dedicating substantial time to polishing these nonverbal behaviors, however, is unlikely to improve student understanding. Faculty can do more to enhance actual learning by skipping the media training sessions and devoting that time directly to their students or class materials.

The price of our current system of student evaluations, moreover, is much higher than the opportunity cost of time spent trying to change nonverbal behaviors. As the previous sections explain, these evaluations impose serious risks of bias. Some of those distortions specifically burden white women, faculty of color, and other traditionally disadvantaged groups. Others penalize any faculty member with the wrong type of face, voice, gestures, or other nonverbal behaviors. The correlation between conventional teaching evaluations and student learning would have to be very high to justify this kind of unfairness.¹⁶¹ Instead, the relationship to learning outcomes is minimal at best.

III. The Cognitive Foundations of Judgment

The psychology literature paints a disturbingly negative picture of student evaluations: These assessments respond primarily to minor aspects of a professor's classroom style; many of

¹⁶¹ Evaluations are most problematic when they are used to determine a particular faculty member's fitness for tenure, promotion, salary increase, or other reward. Even a relatively large correlation between two variables in a population tells us very little about any one population member. A correlation of .50 between student learning and evaluations, for example, is high under social science conventions but means that evaluation ratings explain only 25% of the variance in student learning. *See supra* note 51. Fully three-quarters of the variation in student learning, under these circumstances, would stem from other factors. And despite the overall correlation, some faculty with low evaluations would generate more learning in their students than professors with higher evaluations. For this reason, overall correlations provide a poor basis for judging population members individually. One observer argues that a correlation between two variables should reach .90, a very high threshold, before it is used to rate individuals. Hopkins, *supra* note 51.

those behaviors reflect characteristics like race, gender, and class; and evaluation scores bear little, if any, relationship to student learning.¹⁶² Why are student evaluations so biased?

The problem does not lie with students; it inheres in the process we use to gather their input. Despite some claims to the contrary, students have essential feedback to offer faculty on teaching. They can tell professors what they learned from a course and how that compared to what they expected to learn. They can describe the educational techniques that worked for them and those that did not. They can provide suggestions for how a faculty member might teach differently. Law students can assess the quality of their educational experience in myriad ways.

Obtaining useful feedback from any evaluator, however, requires collecting information under circumstances that allow for meaningful, deliberative reflection. As the Nobel prize-winning psychologist Daniel Kahneman explains, there are two types of human thought processes: System 1 and System 2.¹⁶³ System 1 processes are “fast, automatic, effortless, associative, implicit (not available to introspection), and often emotionally charged; they are also governed by habit and are therefore difficult to control or modify.”¹⁶⁴ System 2 processes are “slower, serial, effortful, more likely to be consciously monitored and deliberately controlled; they are also relatively flexible and potentially rule governed.”¹⁶⁵

As rational adults, we assume that most of our judgments derive from System 2; we perceive ourselves as reasoned creatures who “think things through.” In fact, however, almost all

¹⁶² The extensive literature on teaching evaluations raises other questions on the validity of these measures, particularly their relationship to students’ expected and actual grades. *See supra* note 9. Some of these issues are important to legal education, but are beyond the scope of this article.

¹⁶³ Daniel Kahneman, *A Perspective on Judgment and Choice: Mapping Bounded Rationality*, 58 *American Psychologist* 697, 698 (2003). Kahneman credits Stanovich and West with proposing the terms “System 1” and “System 2,” although the framework rests on work done by Kahneman and his collaborator Amos Tversky. *Id.* Other scholars have used terms like “experiential” and “analytic,” or “experiential” and “rational.” *See, e.g.*, Paul Slovic et al., *Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics*, 31 *J. Socio-Economics* 329 (2002); S. Epstein, *Integration of the Cognitive and Psychodynamic Unconscious*, 49 *American Psych.* 709 (1994).

¹⁶⁴ Kahneman, *supra* note 163, at 698.

¹⁶⁵ *Id.*

judgments begin as System 1 intuitions; we have gut reactions to just about everything.¹⁶⁶ System 2 deliberations challenge and override our initial System 1 decisions in a surprisingly limited number of those cases.¹⁶⁷ Instead, we are more likely to use System 2 cognitive processes to justify the intuitive judgments we generate using the more emotional leaps of System 1.¹⁶⁸

The evaluations that students offer of faculty begin, like most judgments, as System 1 processes. As such, they are heavily influenced by first impressions—even though the students have known the faculty member for at least a semester. When the brain makes a System 1 decision, it does not assess all of the information gathered on a subject over time; that is a System 2 process. Instead, the mind uses its quick-fix, System 1 toolbox: It seizes its first impressions of the subject. In a sense, the students' minds reach back to the first impressions they formed of the instructor, using those as the basis for their evaluation. Those impressions, of course, incorporate all of the biases described in the previous section.

The dominance of System 1 thinking offers a possible explanation for the particularly high correlation between end-of-semester teaching evaluations and “thin slice” judgments of the same teachers. Students' end-of-semester assessments may not just correspond to those first impressions; they may *be* those impressions. Because the process we use to collect end-of-

¹⁶⁶ *Id.* at 716-17.

¹⁶⁷ Although it is impossible to know precisely how often System 2 overrides System 1 in daily life, Kahneman estimates that the “most common” paths of decisionmaking are those that involve only System 1 (i.e., System 2 offers no corrections at all) or those in which System 2 only partially and inadequately corrects System 1. *Id.* at 717.

All “judgments,” in the common meaning of that word, do activate System 2 to some extent, because they require explicit choices. *Id.* at 699. As Kahneman and others have made clear, however, many judgments occur without System 2 overriding the instinctive direction proposed by System 1; they are “intuitive” judgments. *Id.*

¹⁶⁸ As R.B. Zajonc, a pioneer in this area of psychology observed: “We sometimes delude ourselves that we proceed in a rational manner and weigh all the pros and cons of the various alternatives. But this is probably seldom the case Most of the time, information collected about alternatives serves us less for making a decision than for justifying it afterward.” R.B. Zajonc, *Feeling and Thinking: Preferences Need No Inferences*, 35 *American Psychologist* 151, 155 (1980).

semester evaluations does not encourage System 2, deliberative reflection, the students' brains most likely retrieve their first impressions and report those as their current evaluation.¹⁶⁹

Obtaining more meaningful evaluations from students, therefore, requires finding ways to engage System 2 thinking. There are at least three reasons why the current evaluation system fails to encourage that deeper thought. First, System 1 works with special efficiency in deciding whether things are good or bad.¹⁷⁰ Research, in fact, suggests that the brain has “specialized neural circuitry” that renders these good/bad distinctions promptly and confidently¹⁷¹ When the brain is asked to make any good/bad judgment, it is less likely to invoke System 2's more reflective thinking because System 1 handles these decisions so efficiently. This doesn't mean that the brain invariably makes correct good/bad decisions using System 1, but it does tend to rely upon that process for these decisions.

Rating a professor's teaching requires the brain to make good/bad distinctions: We ask students to indicate how “good” or “bad” various aspects of that teaching were. This is especially true of the numerical judgments that most evaluation processes demand from students. The very nature of these forms, which ask students to rate value without requiring other types of reflection, discourages the brain from moving beyond System 1, instinctive reactions.¹⁷²

Second, System 2 thinking requires time. Our brains excel at intuitions, first impressions, and quick judgments, but struggle to make more complex decisions. As one pair of psychologists

¹⁶⁹ Other mechanisms may also contribute to the high correlation between thin slice evaluations of teaching and end-of-semester evaluations. Different theorists have suggested that first impressions may be accurate in this context; that those impressions may be self-reinforcing; and that students may be incapable of forming deeper, more considered assessments of teaching. *See generally* Ambady & Rosenthal, *supra* note 45; Clayson & Sheffet, *supra* note 6. The simplest explanation, however, may be the best: The mind draws upon first impressions as part of its System 1 thinking.

¹⁷⁰ *See* Kahneman, *supra* note 163, at 701; Zajonc, *supra* note 168.

¹⁷¹ Kahneman, *supra* note 163, at 701 (citing Joseph E. LeDoux, *Emotion Circuits in the Brain*, 23 *Annual Rev. Neurosci.* 155 (2000)).

¹⁷² Many law school forms, paralleling those used in other university departments, ask students both to make numeric ratings and to offer comments. The forms, however, almost invariably seek numeric ratings before comments, channeling the brain into fairly simplistic “good/bad” thinking. The number of students who offer comments, moreover, almost always falls far short of those providing numerical scores.

recently observed, “human cognitive capacities are limited” so “[p]roblem solving is hard work.”¹⁷³ Humans have restricted working memory, making it difficult for us to process multiple pieces of data.¹⁷⁴ To make thoughtful evaluations, the brain needs time to recall diverse bits of data, compare them, group pieces of information into larger chunks of partial judgments, and ultimately yield a reasoned response. Researchers have repeatedly established that time pressure induces reflexive, System 1 thinking rather than more reflective System 2 processes.¹⁷⁵

For students, evaluating a professor’s teaching requires more complex deliberation than most faculty members realize. As instructors, our beliefs about teaching process are omnipresent; they shape our class preparation as well as our actions in the classroom on a regular basis. Even if we don’t focus consciously on how we teach, the process is salient to us. Students, however, concentrate on the product of our teaching rather than the process. They attend daily to what they are learning and what they still need to know; the underlying process is secondary. For students to evaluate a semester’s worth of teaching, they must recall the assignments, lectures, explanations, and other characteristics of a professor’s work; examine those memories from a new perspective;

¹⁷³ Klaus Oberauer & Reinhold Kliegl, *A Formal Model of Capacity Limits in Working Memory*, 55 *J. Memory & Language* 601, 601 (2006).

¹⁷⁴ Some psychologists have concluded that the human brain can hold only four “chunks” of data in working memory at one time. See Nelson Cowan, *The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity*, 24 *Behavioral & Brain Sciences* 87, 88 (2000) (reviewing evidence). Other psychologists conceptualize the limits on working memory in other ways. *Id.* at 88 (noting seven alternative views); Oberauer & Kliegl, *supra* note 173 (proposing a particularly sophisticated model). All, however, agree that working memory is constrained.

¹⁷⁵ See, e.g., Anton J. Dijker & Willem Koomen, *Stereotyping and Attitudinal Effects Under Time Pressure*, 26 *Eur. J. Soc. Psych.* 61 (1996) (stereotypes exerted greater role on decisions made under time pressure); Melissa L. Finucane et al., *The Affect Heuristic in Judgments of Risks and Benefits*, 13 *J. Behavioral Decision Making* 1 (2000) (study one) (subjects perceived greater inverse relationship between risks and benefits of activities when considering them under time pressure than when assessing them without time pressure); David M. Sanbonmatsu & Russell H. Fazio, *The Role of Attitudes in Memory-Based Decision Making*, 59 *J. Personality & Soc. Psychol.* 614 (1990) (impressions affect consumer decisions under time pressure). See generally Kahneman, *supra* note 163, at 711.

and judge their efficacy in achieving a variety of goals. Doing this well takes more than five, ten, or even fifteen minutes.¹⁷⁶

Finally, since the human brain has limited capacity, any cognitive distraction impairs System 2 thinking. Laboratory research demonstrates that divided attention reduces reasoned thought, promoting reliance on intuitive, stereotyped, and other automatic System 1 processes.¹⁷⁷ People have trouble thinking about two issues at once or switching quickly from one mental task to another.

Students complete traditional teaching evaluations under tremendous cognitive load. They have just finished a challenging law school class; they may be filling gaps in their notes or digesting the professor's final comments.¹⁷⁸ With the end of the semester near, they probably are worried about their performance in this class or others. They may be thinking about an upcoming job interview, a meeting with a friend, or a reunion with family members; the mind offers a large number of distractions. Laboratory research suggests that even very small distractions—such as memorizing a nine-digit number between tasks—can significantly reduce reflective thought and enhance reliance on stereotypes.¹⁷⁹ The cognitive demands that law students face are much more substantial than these distractions, virtually forcing them to rely on intuitive System 1 channels to complete teaching evaluations.

¹⁷⁶ Although many professors allot 10-15 minutes of class time for students to complete evaluations, most faculty observe that students rarely use all of this time. Law students are anxious to move onto other tasks—lunch, another class, reviewing their notes, a meeting with a friend—and tend to complete these evaluation forms quickly.

¹⁷⁷ See, e.g., C. Neil Macrae et al., *Creating Memory Illusions: Expectancy-Based Processing and the Generation of False Memories*, 10 *Memory* 63 (2002); Sabine Szcesny & Ulrich Kuehnen, *Meta-Cognition About Biological Sex and Gender-Stereotypic Physical Appearance: Consequences for the Assessment of Leadership Competence*, 30 *Personality & Soc. Psychol. Bull.* 13 (2004). See generally Kahneman, *supra* note 163, at 711.

¹⁷⁸ Conversely, if a professor distributes evaluations at the beginning of a class period, students may initially be focused on their previous class, as well as the material they have come to discuss. They must put all of these thoughts aside to focus on evaluation.

¹⁷⁹ Szcesny & Kuehnen, *supra* note 177.

The key to more meaningful evaluations of teaching, in sum, lies in creating the circumstances that allow students to engage more reflective System 2 thought processes before providing those insights. Students need time and mental space to transcend intuitive System 1 judgments; they need mental prompts to help them review the semester's work; and, like all decision makers, they need assistance in moving beyond reflexive "good/bad" judgments. The final section of this article explores how law schools might design such a system.

IV. Deeper Thinking, Better Evaluation

The legal system, like all social structures, rests on judgments that individuals make about others. Lawyers decide which law graduates to hire and promote; clients choose which firms to retain. Negotiators appraise their opponents to calculate a successful offer; deal makers judge how far they can press an advantage. Trial lawyers size up jurors, while jurors assess lawyers, witnesses, and parties. Appellate advocates tailor their arguments to the bench, while the judges register the advocate's sincerity. Student evaluations of teaching are just one type of interpersonal judgment occurring within a constant stream of those judgments.

All of these social evaluations risk the biases described above. Humans rest their judgments of others on intuitive System 1 thinking; nonverbal behaviors, filtered by social stereotypes, powerfully affect those assessments. Designing a process to promote more reflective student evaluations of teaching, therefore, can draw upon research that has identified aids to deliberative decision making in other contexts. Conversely, reducing bias in student evaluations of teaching may illuminate paths to better decision making in other aspects of the legal profession.

At a minimum, thoughtful evaluation of teaching requires time and attention; otherwise, System 2 thought processes fail to engage. Psychologists have identified a number of other

conditions that promote more reflective, deliberative judgments. These include (1) encouraging decision makers to be as accurate as possible;¹⁸⁰ (2) focusing evaluators on the individuality of the person they are assessing;¹⁸¹ (3) reminding decision makers to consider relevant data;¹⁸² (4) facilitating group discussion of judgments;¹⁸³ and (5) establishing accountability to a third party.¹⁸⁴

Our current system of gathering student evaluations incorporates few, if any, of these features. It is possible, however, to identify workable processes that embrace these features. Gregory Munro, for example, recommends using Small-Group Instructional Diagnosis (SGID) to assess law school teaching.¹⁸⁵ In this process, a facilitator meets with small groups of students to gather their impressions of a course and instructor. The students discuss their perspectives as a group, expanding the information available to each student, checking individual biases, establishing accountability, and implicitly noting the seriousness of the process and need for accuracy.¹⁸⁶ The group discussions provide adequate time for thoughtful discussion and reduce cognitive overload by focusing attention on assessment.

For the best results, the facilitator meets with the faculty member before these group discussions, obtaining information about the course goals, the subject matter, and the professor's

¹⁸⁰ Mary E. Wheeler & Susan T. Fiske, *Controlling Racial Prejudice: Social-Cognitive Goals Affect Amygdala and Stereotype Activation*, 16 *Psychol. Sci.* 56, 57 (2005).

¹⁸¹ Wheeler & Fiske, *supra* note 180, at 57.

¹⁸² Kahneman, *supra* note 163, at 711-12.

¹⁸³ Richard F. Martell & Keith N. Leavitt, *Reducing the Performance-Cue Bias in Work Behavior Ratings: Can Groups Help?*, 87 *J. Applied Psychol.* 1032, 1033-34 (2002) (reviewing literature).

¹⁸⁴ Thomas E. Ford et al., *The Role of Accountability in Suppressing Managers' Preinterview Bias Against African-American Sales Job Applicants*, 24 *J. Personal Selling & Sales Mgmt.* 113 (2004); Wheeler & Fiske, *supra* note 180, at 57.

¹⁸⁵ Gregory S. Munro, *Outcomes Assessment for Law Schools* 136 (2000). Gerald Hess and Eric Orts have adopted a similar technique to gather ongoing feedback from law students. Hess calls the groups "student advisory teams," while Orts labels them "quality circles." See Gerald F. Hess, *Student Involvement in Improving Law Teaching and Learning*, 67 *UMKC L. Rev.* 343 (1998); Eric W. Orts, *Quality Circles in Law Teaching*, 47 *J. Legal Educ.* 425 (1997). The technique derives from management tools developed by Edward Deming and implemented by Japanese industries and other businesses worldwide. See Hess, *supra*, at 347-50; Orts, *supra*, at 425-26.

¹⁸⁶ See Munro, *supra* note 185, at 137.

pedagogic strategies.¹⁸⁷ The faculty member may provide a brief written statement of his or her teaching objectives, offering students further focus for their discussion. A professor may also note issues that troubled them during the semester, inviting student feedback on those matters.

Students who participate in these small-group discussions applaud them enthusiastically.¹⁸⁸ Indeed, research suggests that students do not like traditional, end-of-semester written evaluations; they much prefer group discussions that promote dialogue with the professor.¹⁸⁹ Group discussions involve students more thoughtfully in assessment and underscore the school's interest in their input. Students have the opportunity to view the purposes of legal education from a perspective different from their own, often enhancing their own educational commitment and learning strategies.

Faculty members also learn more from these discussions than they do from standard teaching evaluations. Thoughtful discussion among students, with knowledge of the professor's pedagogic objectives, produces more detailed and informative feedback. With modest training, law faculty can successfully facilitate these sessions for one another. As they do so, they broaden their personal knowledge of classroom techniques, student responses, and pedagogic successes or failures. The process of facilitation and evaluation may create an ongoing dialogue about teaching among faculty, underscoring their joint commitment to teaching.

Small group discussion of teaching, moreover, can inform broader curricular goals within the law school. Faculty have few opportunities to consider the curriculum as a whole, or even the place of their courses within a student's three-year law school career. Discussions of a particular

¹⁸⁷ *Id.* at 136.

¹⁸⁸ See, e.g., Robert D. Abbott et al., *Satisfaction with Processes of Collecting Student Opinions About Instruction: The Student Perspective*, 82 J. Educ. Psychol. 201 (1990); Hess, *supra* note 185, at 351-52, 355-61; Orts, *supra* note 185, at 425.

¹⁸⁹ See, e.g., Abbott et al., *supra* note 188. That dialogue may not occur directly; especially when evaluating faculty for tenure or promotion, it is better to use another faculty member to facilitate the student discussion. Faculty, however, can respond to student suggestions through their classroom performance, establishing an ongoing discussion about teaching.

course and instructor can include questions about why the students took the course, how they viewed it fitting within their education, how prepared they feel for advanced offerings in the area, and whether they would have preferred a different type of offering. These individual discussions can fuel broader curricular discussions within the school.

This type of faculty review, of course, takes more time than traditional student evaluations. Most institutions would find it difficult to evaluate every course every semester using small group discussions. It is unlikely, however, that schools genuinely need to assess every course and faculty member that frequently.¹⁹⁰ Schools might aim to evaluate each course once every three years, meaning that about one-sixth of the curriculum would undergo full evaluation each semester. Faculty eligible for tenure or promotion could be evaluated somewhat more frequently. The current evaluation system produces a large amount of data every semester, but those data have limited value. Generating smaller amounts of high-quality information undoubtedly would better serve institutional needs.¹⁹¹

Schools, moreover, can require faculty members to continue gathering regular feedback for formative purposes in every course they teach.¹⁹² Some faculty may do this with conventional written evaluations, although schools should explain the drawbacks of these forms. Other professors may use a modified version of the Small-Group Instructional Diagnosis, conducted by the faculty member him or herself during the semester.¹⁹³ Other techniques for assessing one's

¹⁹⁰ Research suggests that even students tire of the number and repetitiveness of the evaluation forms they complete. See, e.g., Abbott et al., *supra* note 188, at 201.

¹⁹¹ Students currently devote up to 75 minutes per semester (15 minutes for each of five courses) to completing faculty evaluations. Assigning each student to one small-group evaluation session would consume the same amount of student time, give students a more satisfying experience, and generate more productive evaluations.

¹⁹² Evaluators distinguish between “formative” and “summative” evaluations. Formative evaluations are used to inform an ongoing process (such as a class), helping adapt the process to the participants’ needs. Summative evaluations are used to evaluate an experience once it has concluded.

¹⁹³ For helpful discussion of using small group discussions in this manner, see Hess, *supra* note 185; Orts, *supra* note 185.

own performance include one-minute papers and informal, teacher-designed evaluation forms.¹⁹⁴ Students particularly like these techniques when they are used mid-semester, with the faculty member responding to student suggestions.¹⁹⁵

Informal evaluations of this nature would allow students to comment regularly on all courses, while giving faculty members useful feedback in time to implement changes. Indeed, students would gain more from these formative evaluations, which allow a faculty member to respond and adapt, than from the current overload of summative assessments that are used primarily to rate faculty after a course has finished. The combination of individually administered, formative evaluations, used to inform progress and modify a professor's teaching, with small-group assessments in selected classes each semester, gives students the broadest opportunity to provide feedback on instruction and benefit from that process.¹⁹⁶

These student-centered methods of instructional evaluation can also fit within a broader framework that includes teaching assessments offered by colleagues, alumni, education experts, and professors themselves.¹⁹⁷ Evaluation scholars repeatedly stress the need to complement student evaluations with other forms of assessment. Yet traditional student evaluations, which are so easily implemented and generate so much apparently "hard" data, predictably overwhelm these other techniques. Shifting the nature of student evaluations will allow faculties to match those assessments more effectively with those offered by other evaluators.

¹⁹⁴ See Hess, *supra* note 185, at 346.

¹⁹⁵ Abbott et al., *supra* note 188.

¹⁹⁶ It is essential, however, that these formative evaluations remain informal, confidential exercises between the professor and students enrolled in the course. Including these assessments in evaluations of faculty for promotion, salary increases, or other purposes would undercut the goals of the small group discussion process.

¹⁹⁷ For discussion of these techniques, see, e.g., Filippa Marullo Anzalone, *It All Begins with You: Improving Law School Learning Through Professional Self-Awareness and Critical Reflection*, 24 Hamline L. Rev. 324 (2001); Laurie A. Babin et al., *Teaching Portfolios: Uses and Development*, 24 J. Marketing Educ. 35 (2002); Daniel Gordon, *Does Law Teaching Have Meaning? Teaching Effectiveness, Gauging Alumni Competence, and the MacCrate Report*, 25 Fordham Urb. L.J. 43 (1997); Gerald F. Hess, *Improving Teaching and Learning in Law School: Faculty Development Research, Principles, and Programs*, 12 Widener L. Rev. 443 (2006); Markovits, *supra* note 7; Melissa J. Marlow, *Blessed Are They Who Teach an Upper-Level Course, for They Shall Earn Higher Student Ratings*, 7 Fl. Coastal L. Rev. 553 (2006).

Given the multiple benefits of replacing conventional student evaluations with more meaningful processes, law faculties should commit any time required to work out the details of small group discussions or other new methods of evaluating teaching.¹⁹⁸ As scholars, we criticize others for adopting “quick fixes” rather expending the time and resources needed for meaningful evaluations and fair processes.¹⁹⁹ We should be willing to apply the same standards to ourselves. Failing to do so is particularly unjust to our minority colleagues, who appear to suffer disproportionately from current evaluation systems.²⁰⁰ It is also unfair to students, who deserve to be heard more effectively on teaching quality while also learning to thrive with instructors representing diverse races, cultures, and backgrounds.

Improving our method of evaluating teaching is not a panacea for eliminating bias and the role of intuitive judgments in that process. Stereotypes affect how we remember information, as well as how we perceive it.²⁰¹ Reasoned decisions remain anchored in first judgments, giving those initial impressions lasting power.²⁰² Stereotypes based on race, gender, and class are persistent, resisting efforts to overcome them. A professor’s reputation, shaped in part by

¹⁹⁸ The way has been cleared considerably by law faculty and other scholars who have already devoted considerable attention to these details. See, e.g., Hess, *supra* note 185; Orts, *supra* note 185.

¹⁹⁹ A recent search of the Lexis “U.S. and Canadian Law Reviews” file generated 85 articles during the last year alone referring to a “quick fix.” For some of the many “quick fixes” that legal scholars have criticized, see Christina Bohannon, *Reclaiming Copyright*, 23 *Cardozo Arts & Ent. L.J.* 567 (2006); Weitseng Chen, *WTO: Time’s Up for Chinese Banks—China’s Banking Reform and Nonperforming Loan Disposal*, 7 *Chi. J. Int’l L.* 239 (2006); Brian J. Gorman, *Biosecurity and Secrecy Policy: Problems, Theory, and a Call for Executive Action*, 2 *I/S J.L. & Policy Info. Soc’y* 53 (2005-2006); Grant T. Harris, *The Era of Multilateral Occupation*, 24 *Berkeley J. Int’l L.* 1 (2006); A. John Radsan, *Second-Guessing the Spymasters with a Judicial Role in Espionage Deals*, 91 *Iowa L. Rev.* 1259 (2006).

²⁰⁰ See *supra* notes 1-3 and accompanying text. One of the most disheartening aspects of the extensive research and scholarly commentary on teaching evaluations is the almost complete absence of any studies focusing on the influence of race on these evaluations. Katherine Grace Hendrix, who conducted one of the few exploratory studies in this area, commented: “[R]esearchers have overlooked the classroom experiences of teachers and professors of color. In particular, the experience of being a member of a subordinate minority functioning as a professional within a predominantly White educational environment has escaped the interest of the White social scientist.” Katherine Grace Hendrix, *Student Perceptions of the Influence of Race on Professor Credibility*, 28 *J. Black Studies* 738, 739 (1998). See also Huston, *supra* note 8, at 600-01 (commenting on scarcity of research related to racial biases in student evaluations). Law schools have an opportunity to lead the rest of the academy in identifying and remedying these biases.

²⁰¹ See, e.g., Justin D. Levinson, *supra* note 119.

²⁰² See, e.g., Kahneman, *supra* note 163.

stereotypes, may influence even reasoned discussions.²⁰³ And some influences on teaching evaluations, such as students' tendency to downgrade faculty who have graded them negatively,²⁰⁴ fall outside the scope of the biases discussed in this article. Progress in any field, however, requires taking small steps. If we don't attempt to improve the quality of our decision making, we remain trapped forever its lowest levels.

V. Conclusion

Law and legal education assume reflective, rational decision making. Yet psychologists have shown that most of our judgments originate with intuitive preferences. The human brain reacts automatically to nonverbal behaviors and other subtle cues in the environment; social stereotypes further shape our perceptions on an unconscious level. Creating the conditions to support deliberative, reflective thinking is much harder than we believe.

As educators, we can take an important step toward understanding the interplay of intuitive and analytic thinking by examining those processes in the context of routine teaching evaluations. Those assessments draw heavily on the brain's automatic processes, allowing minor stylistic mannerisms and stereotypes to color ratings. Designing evaluation systems that prompt more reflective, rational input would accord students enhanced respect, improve instruction, and treat faculty colleagues more fairly. Exploring such systems will also increase our own understanding of the intricate processes that drive decision making, knowledge we can apply to almost every field of law.

²⁰³ Very little research has examined the influence of instructor reputation on current evaluation systems, but the few studies on this relationship find a very strong effect. See, e.g., Bryan W. Griffin, *Instructor Reputation and Student Ratings of Instruction*, 26 Contemp. Educ. Psychol. 534 (2001) (instructor reputation correlated significantly with end-of-semester ratings, even after controlling for numerous other factors). It is possible that this effect would shape other modes of evaluation as well.

²⁰⁴ See *supra* note 9.