

Indiana Law Journal Fall, 1993

Symposium: Improving Communications in the Courtroom

CALIBRATING THE SCALES OF JUSTICE: STUDYING JUDGES' BEHAVIOR IN BENCH TRIALS

Peter David Blanck[‡]

Copyright © 1993 by the Trustees of Indiana University; Peter David Blanck

Table of Contents

INTRODUCTION	1120
I. A MODEL FOR THE STUDY OF JUDGES' AND JURIES' BEHAVIOR	1126
II. THE STANFORD STUDY: TESTING THE MODEL IN JURY TRIALS	1132
<i>A. Global and Micro Dimensions of Judges' Behavior</i>	1133
<i>B. Simple Relationships in the Model</i>	1136
<i>C. Cumulative Relationships in the Model</i>	1142
III. THE IOWA STUDY: TESTING THE MODEL IN BENCH TRIALS.....	1143
<i>A. Research Strategy and Design: Trials and Participants</i>	1143
<i>B. Research Strategy and Design: On-Line Ratings of the Trials</i>	1144
IV. THE IOWA STUDY: FINDINGS AND IMPLICATIONS	1146
<i>A. Simple Relationships in the Model</i>	1146
<i>B. Cumulative Relationships in the Model</i>	1154
V. EMERGING ISSUES IN THE STUDY OF TRIAL PROCESSES	1168
<i>A. Toward a More Cumulative View of Social Science Research</i>	1168
<i>B. Using the Model to Assess Trial Error</i>	1172
<i>C. Research on Sentencing</i>	1178
CONCLUSION	1180
APPENDICES	1182
<i>A. One View of Nonverbal Behavior in the Courtroom</i>	1182
<i>B. Summary of Data Collection Instruments for Iowa Study of BenchTrials</i>	1183
<i>C. Psychometric Findings</i>	1192

[‡] Professor of Law, University of Iowa College of Law; J.D., 1986, Stanford Law School; Ph.D., 1982, Harvard University. Senior Fellow, The Annenberg Washington Program. This research was supported in part by a grant from the University of Iowa College of Law Foundation and by The Annenberg Washington Program. The Article was strengthened greatly by the input of my colleague David Baldus.

INTRODUCTION

In a series of empirical studies, my colleagues and I have explored judges' and juries' behavior and the "appearance of justice" in actual trials.¹ We have examined various legal and extralegal influences on trial decision-making processes.² In particular, empirical studies we conducted in the California courts focused on the role of judges' behavior and the "appearance of justice" in predicting the outcome of criminal jury trials.³

This Article extends the prior research in several important ways. First, the relationships among legal factors (e.g., evidence) and extralegal factors (e.g., preconceived biases and behavior related to the appearance of justice) is explored through further empirical testing of a theoretical model of courtroom dynamics. These relationships are examined in a study conducted in the Iowa courts, the first such study of extralegal influence in bench trials.

Second, this Article begins to model the combined and independent impact of evidentiary and extralegal factors on decision making and sentencing in criminal bench trials. The analyses test the conclusion by others that, in some cases, extralegal factors such as judges' nonverbal communications to trial participants have a relatively greater impact on trial outcomes than does the strength of the evidence.⁴ The analyses also explore the relation between judges' fact-finding behavior and their subsequent sentencing patterns in bench trials.

Third, this Article extends the model and empirical framework to provide a more comprehensive

¹ See generally Peter D. **Blanck**, What Empirical Research Tells Us: Studying Judges' and Juries' Behavior, 40 AM. U. L. REV. 775 (1991) [hereinafter Empirical Research]; Peter D. **Blanck** et al., The Measure of the Judge: An Empirically-Based Framework for Exploring Trial Judges' Behavior, 75 IOWA L. REV. 653, 655-57 (1990) [hereinafter The Measure of the Judge] (describing research that studied judges' behavior to determine if it "appears" to trial participants to be fair and impartial); Peter D. **Blanck** et al., Note, The Appearance of Justice: Judges' Verbal and Nonverbal Behavior in Criminal Jury Trials, 38 STAN. L. REV. 89, 89-97 (1985) [hereinafter The Appearance of Justice] (discussing judicial influence and its relationship to procedural due process).

As I noted in an earlier study, due process in a criminal trial mandates that the trial judge not display bias toward the defendant. Empirical Research, *supra*, at 776.

Trial judges are not only required to be fair and impartial, they must also "satisfy the appearance of justice." The judge's appearance or extralegal behavior must not indicate to the jury or to other trial participants that the judge believes the defendant is guilty. The appearance of judicial bias alone is grounds for reversal even if the trial judge is completely impartial. *Id.* at 776 (footnotes omitted); see *id.* at 776 nn.1-3 and cases cited therein; see also *Etzel v. Rosenbloom*, 189 P.2d 848, 852 (Cal. Dist. Ct. App. 1948) ("Four things belong to a judge: to hear courteously, to answer wisely, to consider soberly, and to decide impartially." (quoting Socrates)).

² Other researchers have argued that the information available to jurors in criminal trials falls into two primary categories: evidentiary (e.g., pertaining to the crime) and nonevidentiary (e.g., pertaining to defendants' background characteristics). See Martin F. Kaplan & Gwen DeArment Kemmerick, Juror Judgment as Information Integration: Combining Evidential and Nonevidential Information, 30 J. PERSONALITY & SOC. PSYCHOL. 493 (1974).

³ See The Appearance of Justice, *supra* note 1 (reporting results of study of 34 misdemeanor jury trials conducted in California Municipal Court); Communicating with Juries, Panel One: Judge-Jury Communications: Improving Communications and Understanding Bias, The Annenberg Washington Program Conference, April 10, 1992, 68 IND. L.J. 1037 (1993) [hereinafter Panel One] (introductory remarks of moderator Steven Adler) (suggesting that there is not necessarily one key moment during trial that influences juries' decision-making processes; rather, there are complex relationships among courtroom players).

⁴ For a review, see J. Alexander Tanford & Sarah Tanford, Better Trials Through Science: A Defense of Psychologist-Lawyer Collaboration, 66 N.C. L. REV. 741 (1988) (responding to prior criticism that availability of prior psychological research on trial processes enhances lawyers' abilities to influence jurors).

view of the nature of the adversarial process.⁵ For instance, it examines the impact of the perceived competence and influence of the trial participants-judge and counsel-on trial outcomes and sentencing. Moreover, this Article looks at courtroom behavior in lower state courts, where it is estimated that ninety to ninety-five percent of all cases are handled.⁶ Further study of the adversarial process in this context is needed because "next to the police, the lower criminal courts play the most important role in forming citizen impressions of the American system of criminal justice."⁷ In these ways, this Article sets forth a method for researchers, practitioners, and courts to assess factors that may impact, sometimes impermissibly, on decision making in actual trials.

Part I of this Article describes the theoretical model and research framework, including the various legal and extralegal measures for exploring the impact of judges' behavior on trial outcomes and sentencing in bench trials.⁸ Part II summarizes findings from a prior test of the model in California jury trials (the "Stanford Study"). Parts III and IV then present the method and results for the test of the model in Iowa bench trials (the "Iowa Study"). Finally, Part V discusses emerging issues in the analysis of courtroom behavior, including the study of trial error, disqualification law, and sentencing patterns.

Judges, like all human beings, develop certain beliefs about the defendant's guilt or innocence.⁹ Courts, legal scholars, practitioners, and social scientists recognize that extralegal influences may have important effects on trial processes and outcomes.¹⁰ Some trial and appellate courts acknowledge that juries, witnesses, or other trial participants accord great weight and deference to even the most subtle behaviors of the judge.¹¹ Little information, however, is available about the extent to which trial judges themselves are sensitive to, or even conscious of, the effect that

⁵ See *id.* at 759-71 (arguing that critics of psychology and law research misunderstand the complexity of trial processes). For a review of the importance of replication of social science research, see Robert Rosenthal, *How Are We Doing in Soft Psychology?*, 45 *AM. PSYCHOLOGIST* 775 (1990) (demonstrating the importance of replication and cumulation of social science research).

⁶ John Hagan & Kristin Bumiller, *Making Sense of Sentencing: A Review and Critique of Sentencing Research*, in *RESEARCH ON SENTENCING: THE SEARCH FOR REFORM II* (Alfred Blumstein et al. eds., 1983) [hereinafter *RESEARCH ON SENTENCING*].

⁷ MALCOLM M. FEELEY, *THE PROCESS IS THE PUNISHMENT: HANDLING CASES IN A LOWER CRIMINAL COURT* xv (1979).

⁸ See *infra* part I (discussing model to describe and document judge behavior).

⁹ See *The Appearance of Justice*, *supra* note 1, at 90 (noting that the judge may reveal these beliefs during trial by directing the trial based on the judge's own expectations); see also Panel One, *supra* note 3, at 1040 (statement of Steven Adler) (observing that the entire trial sends a message to jurors about who is right and wrong); *id.* at 1041 (statement of Robert Rosenthal) (noting that judges' expectations may sometimes act as a self-fulfilling prophecy); RICHARD J. BERNSTEIN, *BEYOND OBJECTIVISM AND RELATIVISM* 129 (1983) (arguing that human beings can never be devoid of prejudices) (citing HANS-GEORG GADAMER, *PHILOSOPHICAL HERMENEUTICS* 9 (David E. Linge trans., 1976)).

¹⁰ See Panel One, *supra* note 3, at 1040 (statement of Judge LaDoris Cordell) (noting that Stanford Study was undertaken by a coalition of academicians, social scientists, practitioners, and judges); see also HARRY KALVEN, JR. & HANS ZEISEL, *THE AMERICAN JURY* (1966) (providing classic study of judges and juries); JOHN P. RYAN ET AL., *AMERICAN TRIAL JUDGES: THEIR WORK STYLES AND PERFORMANCE* (1980) (encompassing comprehensive analysis of judges' behavior); *The Measure of the Judge*, *supra* note 1, at 654 & n.12 (discussing books, articles, and studies recognizing importance of judge's behavior).

¹¹ As Judge Jochems remarked in 1930, "[t]he trial judge occupies a high position. He presides over the trial. The jury has great respect for him. They can be easily influenced by the slightest suggestion coming from the court, whether it be a nod of the head, a smile, a frown, or a spoken word." *State v. Wheat*, 292 P. 793, 797 (Kan. 1930) (Jochems, J., dissenting), quoted in *State v. Hamilton*, 731 P.2d 863, 868 (Kan. 1987); *Marino v. Cocuzza*, 81 A.2d 181, 185 (N.J. Super. Ct. App. Div. 1951); see also *The Appearance of Justice*, *supra* note 1, at 155-56 (giving pattern jury instruction warning that behavior of judge during trial should not influence jury decision making).

their extralegal behavior might have on fact finding, trial outcomes, or their sentencing patterns.¹² The few studies conducted indicate strong judicial interest in exploring the connection between courtroom behavior and trial outcomes.¹³

Existing judicial training programs across the country explore the importance of judges' behavior and decision making in the courtroom.¹⁴ Most trial judges, however, receive little feedback about their courtroom communication, and what little they do receive is mostly anecdotal.¹⁵ This may be in part because there are few standardized methods through which such feedback may be provided,¹⁶ judges are reluctant to receive such feedback,¹⁷ or judges lack effective techniques for monitoring the impact of their courtroom behavior.¹⁸

Appellate courts assess the propriety and impact of legal and extralegal factors on jury and judge decision making by balancing a number of factors.¹⁹ These factors include the relevance and nature of the behavior, the efficiency of any instruction used to cure the error, and the prejudicial effect of the behavior in light of the entire trial.²⁰ However, the validity of these factors has not been tested through empirical study of actual trials, and the lack of systematic information is glaring.²¹ Moreover, ad hoc, case-by-case determinations of the appearance of justice and trial error result because reviewing courts typically evaluate the record as a whole in determining trial fairness, judicial bias, or prejudicial error.²² The present research continues the

¹² See, e.g., Martha Minow, Stripped Down Like a Runner or Enriched by Experience: Bias and Impartiality of Judges and Jurors, 33 WM. & MARY L. REV. 1201, 1203 (1992) (questioning whether judges or jurors "know bias when they see it"); see also LaDoris H. Cordell & Florence O. Keller, Pay No Attention to the Woman Behind the Bench: Musings of a Trial Court Judge, 68 IND. L.J. 1199 (1993) (arguing that jurors attribute certain opinions, feelings, and biases to judicial personnel through transference).

¹³ See Panel One, supra note 3, at 1043 (statement of Robert Rosenthal) (expressing concern about judges' behavior and its effect on juries); see also George Everson, The Human Element in Justice, 10 J. CRIM. L. & CRIMINOLOGY 90, 99 (1919) (providing an early study of magistrates, which concluded that "[t]hese studies of the work of magistrates' records ... are startling because they show us so clearly to how great an extent justice resolves itself into the personality of the judge"); cf. CATHERINE FITZMAURICE & KEN PEASE, THE PSYCHOLOGY OF JUDICIAL SENTENCING 7 (1986) (providing a comparative study of sentencing and noting alleged remark by Lord Chief Justice that research on judicial behavior "would not tell judges anything they did not already know ..."); Robert Hanley, A Courtroom Experiment in High-Tech Video, N.Y. TIMES, Nov. 26, 1992, at B8 (discussing judge's informal study of videotapes of courtroom proceedings to analyze his own performance and behaviors).

¹⁴ The Measure of the Judge, supra note 1, at 676 (discussing program that videotapes and analyzes judge's behavior during trial proceedings).

¹⁵ Panel One, supra note 3, at 1046 (statement of Judge LaDoris Cordell); Cordell & Keller, supra note 12, at 1202-03.

¹⁶ Panel One, supra note 3, at 1045-46 (statement of Judge LaDoris Cordell) (suggesting that jurors, at the end of a trial, are excellent sources of feedback for trial judges about individual judge's performance).

¹⁷ In this regard, it may be interesting to compare the behavior of elected versus appointed state court judges.

¹⁸ Panel One, supra note 3, at 1053 (statement of Robert Rosenthal) (commenting that communication trainers have not yet proven to be effective).

¹⁹ See *United States v. Olgin*, 745 F.2d 263, 268-69 (3d Cir. 1984) (emphasizing that "[t]he reviewing court should be more concerned with a comment on a matter central to the defense than with comment on a tangential issue"), cert. denied, 471 U.S. 1099 (1985); *United States v. Anton*, 597 F.2d 371, 374-75 (3d Cir. 1979) (noting that judge's comment on defendant's credibility was one factor in appellate court's reversal of conviction). See generally *infra* part V (discussing error and judges' behavior).

²⁰ *Oglin*, 745 F.2d at 268-69; *The Appearance of Justice*, supra note 1, at 95-96 (reviewing appellate courts' factor approach in assessing propriety of judge's behavior). See generally *infra* part V.

²¹ See Charles J. Ogletree, Jr., *Arizona v. Fulminante: The Harm of Applying Harmless Error to Coerced Confessions*, 105 HARV. L. REV. 152, 163 (1991) (noting the lack of systematic study of harmless error rule). See generally Gary B. Melton, *The Law Is a Good Thing (Psychology Is, Too)*, 16 LAW & HUM. BEHAV. 381, 387-88 (1992) (suggesting some positive effects of psychological inquiry in the legal system).

²² Emily Wheeler, *The Constitutional Right to a Trial Before a Neutral Judge: Federalism Tips the Balance Against State Habeas Petitioners*, 51 BROOK. L. REV. 841, 850-51, 875, 876 n.181 (1985) (citing cases that demonstrate this trend).

development of more systematic approaches for assessing trial fairness, as well as the legal parameters of trial error.²³ This program of study is warranted, considering the view by some that the viable role of federal and state appellate courts in overseeing the appearance of justice and the conduct of trial judges is even more important than the right to an impartial judge.²⁴

Legal practitioners are similarly interested in the impact of judges' behavior on courtroom fairness. The American Bar Association's 1990 amendments to the Model Code of Judicial Conduct include a canon that emphasizes the need for the appearance of fairness and justice in the courtroom.²⁵ Yet, like judges, legal practitioners rarely rely on systematic methods for assessing judges' or other trial participants' behavior and its impact on trial fairness.²⁶ To the contrary, practitioners often confuse conceptions of trial fairness and judicial impartiality.²⁷ Professor Leubsdorf writes: "Educated by the Legal Realists and their successors, lawyers fear that the values and experiences of judges ultimately shape their decisions. Yet lawyers also believe that it must mean something to speak of a judge as impartial, and we also suspect that the role of law depends on the belief that the rule of law is more than a masquerade."²⁸

Some empirical studies by social scientists help reveal the complexity of the study of judges' behavior in jury and bench trials²⁹ and replace unsubstantiated myths about courtroom behavior with empirically validated conclusions.³⁰ Prior studies suggest that although trial outcomes are

²³ See *infra* part V (discussing emerging issues in the study of trials).

²⁴ See, e.g., LAURENCE H. TRIBE, *AMERICAN CONSTITUTIONAL LAW* § 2-3, at 17-18, § 3-39, at 147 (1978) (discussing the constitutional role of state courts and the appellate system); Wheeler, *supra* note 22, at 850-51 (describing the appellate courts' responsibility to maintain standards of justice in lower criminal courts, and concluding that the lack of a well-defined constitutional right to an impartial judge means that claims of judicial bias merge with unfair trial claims, which often do not violate the due process standard of fundamental fairness).

²⁵ The Code states:

A judge shall perform judicial duties without bias or prejudice. A judge shall not, in the performance of judicial duties, by words or conduct manifest bias or prejudice ... based upon race, sex, religion, national origin, disability, age, sexual orientation or socioeconomic status, and shall not permit staff, court officials and others subject to the judge's direction and control to do so. MODEL CODE OF JUDICIAL CONDUCT Canon 3(B)(5) (ABA Standing Committee on Ethics and Professional Responsibility 1990).

The commentary to the canon states:

A judge must perform judicial duties impartially and fairly. A judge who manifests bias on any basis in a proceeding impairs the fairness of the proceeding and brings the judiciary into disrepute. Facial expressions and body language, in addition to oral communication, can give to parties or lawyers in the proceeding, jurors, the media and others an appearance of judicial bias. A judge must be alert to avoid behavior that may be perceived as prejudicial.

Id.

²⁶ See Charles-Edward Anderson, *Trial by Press?: Pretrial Publicity Doesn't Bias Jurors*, Panelists Say, A.B.A. J., Sept. 1990, at 32 (reporting consensus of panelists at The Annenberg Washington Program that jurors subject to extensive publicity can put aside preconceptions if judges provide proper instructions and other curative assistance); John B. McConahay et al., *The Uses of Social Science in Trials with Political and Racial Overtones: The Trial of Joan Little*, 41 *LAW & CONTEMP. PROBS.* 205, 213-20 (1977) (recounting juror selection strategies based on empirical model, personality traits, and juror nonverbal behavior); Panel One, *supra* note 3, at 1044 (statement of Charles Rulff) (commenting that lawyers can do little to effect communication problems and bias in the courtroom).

²⁷ See John Leubsdorf, *Theories of Judging and Judge Disqualification*, 62 *N.Y.U. L. REV.* 237, 245 (1987) (suggesting that in this confusion is a repressed crisis of confidence in the idea of judicial impartiality).

²⁸ *Id.* at 245 (emphasis in original). Leubsdorf also concludes that the appearance of justice standard "can best be understood as an unsatisfactory attempt to mediate between introspection and objectivity." *Id.* at 277.

²⁹ See Panel One, *supra* note 3, at 1042-43 (statement of Robert Rosenthal) (describing social science research on self-fulfilling prophecies).

³⁰ See Vladimir J. Konecni & Ebbe B. Ebbesen, *External Validity of Research in Legal Psychology*, 3 *LAW & HUM. BEHAV.* 39, 40-42 (1979) (criticizing reliance on simulated legal research for developing practical recommendations); see also Peter D.

not always the product of rational legal analysis, popular criticisms or jury or judge decisionmaking abilities often are based on perceived problems with the comprehensibility of the law or other procedural factors portrayed as inherent to our system of justice.³¹

This Article describes ongoing work on an empirically based framework and model that tests the observation that extralegal factors, sometimes only the subtle and unintentional nonverbal behavior of judges, might alone predict trial outcomes and sentencing patterns.³² The efforts in the Stanford Study and the Iowa Study are highlighted next as a step toward developing this body of research.³³

I. A MODEL FOR THE STUDY OF JUDGES' AND JURIES' BEHAVIOR

In a criminal trial, a trial judge's beliefs or expectations for a defendant's guilt may be manifested either verbally or nonverbally (by facial gestures, body movements, or tone of voice) and can be reflected in a judge's comments on evidence, responses to witness testimony, reactions to counsels' actions, or in rulings on objections.³⁴ Improper beliefs or expectations, if manifested in a judge's behavior, could warrant reversal and judicial disqualification.³⁵ Prior theories of reversible error and judicial disqualification have not been framed adequately because models of "how judges should behave when they do sit" are lacking.³⁶

The Stanford Study explored the effects of judges' behavior on jury verdicts and on other trial

Blanck & Arthur N. Turner, Gestalt Research: Clinical-Field-Research Approaches to Studying Organizations, in HANDBOOK OF ORGANIZATIONAL BEHAVIOR 109, 111 (Jay W. Lorsch ed., 1987) (stating clinical field research is more appropriate where goal is to improve practice).

³¹ See Panel One, supra note 3, at 1037-39 (statement of Steven Adler) (suggesting that legal procedures, as well as the complexity of the law, contribute to juror misunderstanding); see also Michael J. Saks, Do We Know Anything About the Behavior of the Tort Litigation System-and Why Not?, 140 U. PA. L. REV. 1147 (1992) (reviewing empirical evidence on behavior of tort litigation system and demonstrating the inadequacy of evidence for drawing conclusions about the way the system actually performs); Tanford & Tanford, supra note 4, at 742 (arguing that prior critics of legal system exaggerate the relative importance of legal and extralegal factors on trial outcomes).

³² See infra text accompanying notes 159-63 (discussing research results on effects of judges' nonverbal behavior on trial outcomes). As my colleagues and I have suggested elsewhere, the intention in "modeling" courtroom behavior is not to suggest that there is a bright line standard for detecting, quantifying, or measuring the legally permissible limits of judges' behavior. Measure of the Judge, supra note 1, at 675. Nor is it the intention to suggest that trial judges display great stone faces, showing no emotion or reaction to the events in the courtroom. The goal of this Article is to highlight the ongoing tests and refinements of the model of study to aid in the description of courtroom behavior.

³³ For a discussion of the generalizability of the results of social science research and the importance of replication, see The Appearance of Justice, supra note 1, at 136-37 (discussing Stanford Study implications). See also Tanford & Tanford, supra note 4, at 753 (discussing the importance of replicating social science research in mock and real settings).

³⁴ See The Appearance of Justice, supra note 1, at 90-91. Some of the principal ways in which judges can impermissibly influence the criminal trial process include: (1) disparaging remarks toward the defendant; (2) bias in rulings or comments; (3) consideration of matters not in evidence; (4) forming expectations for trial outcome before the defense has presented its case; (5) inappropriate statements of opinion to the jury during the trial; and (6) failing to control the misconduct of counsel. *Id.*

³⁵ In the William Kennedy Smith rape trial, the prosecution alleged that the trial judge's bias for the defendant was reflected through her nonverbal behavior. The prosecution attempted to disqualify the judge on the grounds, among others, that her facial expressions and body language expressed preconceived biases about her views of the ultimate trial outcome. One view of the prosecution's attempt to disqualify the judge is illustrated in Appendix A. See also Leslee Daugherty, *State v. Fie: Determining the Proper Standard for Recusal of Judges in North Carolina*, 65 N.C. L. REV. 1138, 1139-40 (1987) (arguing that recusal is appropriate if certain factors are met or if nonrecusal would give the "appearance" of impropriety).

³⁶ See Leubsdorf, supra note 27, at 239 (arguing also the converse: that models of judicial behavior cannot be developed without facing their implications for disqualification). In addition, judges may be perceived by others as not understanding the force and impact of their behavior at trial. See supra notes 11-12 and accompanying text.

process variables.³⁷ The preliminary research model identified several types of variables that need to be studied to achieve a systematic understanding of judges' behavior and its potential influence on trial decision making and sentencing patterns.³⁸ The model is illustrated in Figure 1 and elements of the extended version of the model are summarized as follows.³⁹

³⁷ See *The Appearance of Justice*, supra note 1, at 101-37 (proposing preliminary model and describing its characteristics); see also Robert Rosenthal, *Pavlov's Mice, Pfungst's Horse and Pygmalion's PONS: Some Models for the Study of Interpersonal Expectancy Effects*, in *THE CLEVER HANS PHENOMENON* 185-89 (Thomas A. Sebeok & Robert Rosenthal eds., 1981) (discussing general model for study of interpersonal expectancy effects).

³⁸ Early studies of courtroom behavior primarily explored bivariate relationships, such as the relation of race and sentencing. The model here employs bivariate and multivariate techniques, allowing a more comprehensive view of causal relationships in the courtroom. See Hagan & Bumiller, supra note 6, at 2-4.

³⁹ The model is extended to include the strength of the evidence variable ("C"), sentencing patterns ("F"), and the competence/influence variable ("G"). These factors were not tested in the Stanford Study. For earlier descriptions of the model, see *The Appearance of Justice*, supra note 1, at 102 (describing findings for "A-B-C-D" simple relationships only); *The Measure of the Judge*, supra note 1, at 680-84 (highlighting "A-B-C-D-E-F" model without presentation of preliminary empirical findings).

FIGURE 1

GENERAL MODEL FOR THE STUDY OF JUDGES' AND JURIES' BEHAVIOR

<i>Variable</i>					Trial	Judge/Jury		Competence/
<i>Name</i>	Background	Expectancy	Behavior	Evidence	Outcome	Agreement	Sentence	Influence
<i>Relationship[†]</i>	A	B	C	C	D	E	F	G
<i>Examples</i>	Defendant's Age, Gender, Criminal History	Judge's Belief <i>Prior</i> to Verdict	Global Styles & Micro Behaviors	Strength/ Complexity	Judge/Jury Verdict	Judge's Belief <i>After</i> Verdict	Magnitude	Judge, Counsel, Witness

"A": background variables

Background variables refer to attributes of the trial participants, such as age, gender, race, socioeconomic status, intellectual ability, and other personal history factors.⁴⁰

"B": attitudinal or "expectancy" variables

A judge's beliefs or expectations about trial outcome (whether actual or perceived by others to be true) can influence the decisionmaking process of the judge or jurors. When judges expect a certain trial outcome, they intentionally or unintentionally behave in a way that indicates what they think the outcome should be.⁴¹

"C": mediating verbal and nonverbal behaviors

Verbal and nonverbal behaviors communicate judges' beliefs and expectations to the trial

[†] The simple bivariate relationships are between any two variables in the model.

Cumulative or multivariate relationships (e.g., "A-B-C" predicting "D") involve more than two variables.

⁴⁰ See *The Appearance of Justice*, supra note 1, at 105 & nn. 55-57 (reviewing studies demonstrating effects of defendant's race on sentencing in death penalty cases and effects of defendant's criminal history on jury verdicts). In the Stanford Study, we were particularly interested in the extent to which information about defendants' criminal histories might predict, or be predicted by, the other variables in the model. This is because frequently the defendants' prior criminal history is known to the judge but not to the jury, unless the defendant takes the stand to testify, thus making, for example, an "A-C" relationship even more striking. Cf. James W. McElhaney, *The Rub*, A.B.A. J., Dec. 1990, at 80-83 (noting that most important decision for criminal defendant is whether to take the stand and testify on his or her own behalf, and most important factor in making that decision is whether judge is going to let prosecutor cross-examine defendant with regard to prior convictions). The seriousness of the offense, as measured by the maximum possible sentence, or type of legal representation, as measured by defendant's choice between legal counsel or pro se defense, are other examples of possible background variables. See infra notes 109-12 and accompanying text; Steven Garber et al., *The Role of Extralegal Factors in Determining Criminal Case Disposition*, in *RESEARCH ON SENTENCING*, supra note 6, at 129, 132-35, 138 (listing other potential background variables, and noting, for example, that seriousness of the offense is the most important determinant clients use in choosing an attorney).

⁴¹ In doing so, judges set into motion behavior and trial processes that may increase the likelihood of the occurrence of a certain trial outcome. See *The Appearance of Justice*, supra note 1, at 132-36; see also *State v. Armour*, 1991 Ohio App. LEXIS 6098, at §44-§64 (8th Dist. Dec. 19, 1991) (Corrigan, J., concurring & Harper, J., dissenting) (discussing the appearance of justice, judges' expectancy effects, and trial fairness).

Of course, judges and jurors are not always influenced in their decision making by their initial expectations, but neither do they always make decisions based on the evidence alone. See Tanford & Tanford, supra note 4, at 750-51 (noting psychological research showing that jurors have difficulty putting aside their biases to reach accurate verdicts); cf. id. at 748 (noting that those critical of psychology and law research tend to assume that jurors are either unbiased or can easily set aside their biases so that they can decide cases solely on the basis of the evidence). See generally *INTERPERSONAL EXPECTATIONS: THEORY, RESEARCH AND APPLICATION* (Peter D. Blanck ed., 1993) [hereinafter *INTERPERSONAL EXPECTATIONS*] (providing an overview of expectancy effects in applied contexts).

participants.⁴² In the present empirical framework, verbal and nonverbal behaviors are coded on various dimensions using ratings of judges' "global" styles (e.g., judicial or directive style) and "micro" behaviors (e.g., nonverbal behaviors such as eye contact or head nods) toward different trial participants.⁴³

"C": mediating strength and complexity of the evidence

This variable explores whether, in close cases where the evidence may be only marginally important, judges' communicative behavior during bench trials becomes an increasingly important predictor of trial outcomes.⁴⁴ Inclusion of this variable also helps explore the extent to which jurors base their decisions on the strength and complexity of the evidence presented at trial.⁴⁵ Studies of courtroom behavior typically do not include a measure of the strength or quality of the evidence because this information is often costly to compile.⁴⁶

⁴² A judge's facial expressions, tone of voice, or other nonverbal behavior can, by itself, influence jury verdicts, sometimes in impermissible ways. For example, in the often cited case, *State v. Barron*, 465 S.W.2d 523, 527-28 (Mo. Ct. App. 1971), a Missouri appellate court reversed a burglary conviction on the grounds of nonverbal prejudicial error by the judge. When listening to the defendant's brother testify that the defendant was at home watching television when the alleged burglary occurred, the trial judge placed his hands to the sides of his head, shook his head negatively, and leaned back, swiveling his chair 180 degrees. *Id.*

More recently, in analogous situations, social scientists demonstrated that teacher, doctor, and psychotherapist nonverbal behaviors significantly influence the course of social interaction. See *The Appearance of Justice*, supra note 1, at 108-09 & n.67 (providing investigations that study factors affecting expectancy effects and processes of interpersonal communication that transmit effects). See generally Peter D. Blanck & Robert Rosenthal, *Nonverbal Behavior in the Courtroom*, in *APPLICATION OF NONVERBAL BEHAVIORAL THEORIES AND RESEARCH* 89 (Robert Feldman ed., 1992) (summarizing analogous studies and their relevance to model).

⁴³ In the Iowa Study, two observers independently rated and coded the behavior of judges throughout entire trials. The observers also coded the behavior of the other trial participants, as well as other variables in the model. See infra notes 113-14 and accompanying text. Micro nonverbal behaviors are seven discretely coded actions regularly employed in the study of nonverbal behavior, including: (1) amount of eye contact with the jury; (2) number of smiles; (3) number of head nods; (4) number of significant hand movements; (5) number of forward leans to and away from the jury; (6) number of significant changes in body position (fewer shifts designated as "postural attention"); and (7) number of self-touching behaviors, such as chin-rubbing. *INTERPERSONAL EXPECTATIONS*, supra note 41, at 342-53. For detailed descriptions of the research method, see *id.* at 342-53 (discussing development of working model, data-gathering techniques, and calculation of results); *The Appearance of Justice*, supra note 1, at 113-37 (describing research strategy and design and providing detailed analysis of results of judicial influence study); *The Measure of the Judge*, supra note 1, at 667-75 (developing practical framework for describing and assessing judges' behavior).

⁴⁴ See *The Measure of the Judge*, supra note 1, at 684. Likewise, the complexity of the evidence, as rated throughout the trial, may be helpful in predicting other variables in the model and trial outcome; see also, Garber et al., supra note 40, at 144-45 (suggesting that bias in the role of extralegal factors is larger when extralegal factors play a relatively small role in affecting case disposition).

⁴⁵ See Christy A. Visher, *Juror Decision Making: The Importance of Evidence*, 11 *LAW & HUM. BEHAV.* 1-18 (1987) (commenting that in majority of cases studied, jurors base their verdicts more heavily on evidence and law than on extralegal variables). The Stanford Study did not explore the strength and complexity of the evidence because that study focused primarily on the impact of judges' communicative behavior on juries. Likewise, the main focus of *The American Jury* research was on describing (1) the impact of extralegal information in the 25% of the cases in which there was judge/jury disagreement, and (2) how such information accounts for their finding that, in the vast majority of these cases where disagreement occurred, the jury was more lenient than the judge. *KALVEN & ZEISEL*, supra note 10, at 55-65.

⁴⁶ See Garber et al., supra note 40, at 140 (noting that the quality and strength of the evidence plays an important role in the decision to prosecute, the choice of plea, and trial conviction).

The Iowa Study measures the relative strength of the evidence, in terms of evidentiary support for a guilty or not guilty verdict in bench trials. Ratings of the evidence are made at the conclusion of the trial. See infra text accompanying notes 116-18; cf. *JAMES EISENSTEIN & HERBERT JACOB, FELONY JUSTICE: AN ORGANIZATIONAL ANALYSIS OF CRIMINAL COURTS* 182-83 (1977) (differing results between archival study focusing on quality of evidence and present study of ratings of actual evidence). Other researchers have explored the impact of the strength of the evidence on jury decision making. In one study, mock jurors were presented with several independent charges against a defendant. Norbert L. Kerr & Gary W. Sawyers, *Independence of Multiple Verdicts Within a Trial by Mock Jurors*, 10 *REPRESENTATIVE RES. IN SOC. PSYCHOL.* 16 (1979). The strength of the evidence for each charge was varied experimentally. The results show that as the strength of the

"D": outcome variables

The model measures ultimate trial outcome by the judge's (e.g., in a bench trial) or jury's finding of guilt or innocence. Trial outcome is predicted by several combinations of variables in the model.⁴⁷

"E": judge/jury agreement/disagreement variables

This variable assesses the magnitude of the agreement or disagreement between judge and jury regarding trial outcome.⁴⁸ The "E" variable is not tested in the Iowa Study of bench trials.

"F": sentence imposed variable

This variable assesses the magnitude of the sentence for a particular charge.⁴⁹ Judges retain discretion in the sentencing process,⁵⁰ often considering individual and community perceptions of the crime, the background of the criminal, and the circumstances of the case.⁵¹ The model assesses the impact of defendants' background, behavior, and culpability factors on the sentencing process.⁵² Analyses of these relationships allow for assessment of the degree to

evidence toward guilt is increased on one charge, the mock jurors are less likely to find guilt on the other charge. The researchers interpret the results as attempts by the mock jurors to produce verdicts that would be equitable to the defendant.

In another study, mock jurors were presented with either strong or weak evidence in a murder case. Stanley Sue et al., *Effects of Inadmissible Evidence on the Decisions of Simulated Jurors: A Moral Dilemma*, 3 J. APPLIED SOC. PSYCHOL. 345 (1973). These jurors were presented also with additional evidence that was ruled as either admissible or inadmissible. The findings show that jurors in the weak ("close") evidence condition are relatively more biased by the inadmissible evidence in their decision making. Conversely, jurors in the strong evidence condition are not biased by the inadmissible evidence and produce more guilty verdicts. These findings parallel suggestions here that in close or marginal cases, extralegal factors, such as the judges' communicative behavior, independently influence trial outcomes and sentencing. See *infra* notes 94-98 and accompanying text (noting the results for "C-D" chain). The Sue et al. study also found that jurors in the strong evidence condition showed greater confidence in their subsequent verdicts. Cf. *infra* notes 156-58 and accompanying text (regarding "C-C" results).

In sum, the Iowa Study explores the strength and complexity of the evidence and its relationship to other variables in the model. The study analyzes in actual trials what Kalven and Zeisel call the importance of "sentiments" in close cases. KALVEN & ZEISEL, *supra* note 10, at 111-12, 165 (suggesting that closeness of evidence in a case makes it possible for the jury to respond to sentiment or intuitive feelings "by liberating it from the discipline of the evidence" (emphasis in original)). Kalven and Zeisel suggest that jury (or, by implication, judicial) sentiments about the defendant (e.g., empathy with the defendant) or about the law (e.g., fairness of the law) may provide additional insight into the reasons for decision making in close cases.

⁴⁷ See *infra* notes 82-84, 89-90, 93-98, 135, 149-51, 159-61, 170-71 and accompanying text (discussing the results of simple relationships in the model where one variable is trial outcome). In the Stanford Study, for instance, defendants with more severe criminal histories were more likely to be found guilty, while those with less severe criminal histories were more likely to be acquitted. The Appearance of Justice, *supra* note 1, at 129; see *supra* text accompanying note 82.

⁴⁸ In the Stanford Study, this variable assessed the judge's attitude about the trial process and outcome after the jury reached its verdict.

⁴⁹ A large number of empirical studies on sentencing patterns and behavior have been conducted. See LAWRENCE S. WRIGHTSMAN, *PSYCHOLOGY AND THE LEGAL SYSTEM* 372-99 (2d ed. 1991) (reviewing empirical studies). See generally FITZMAURICE & PEASE, *supra* note 13 (reporting comparative study of the psychology of sentencing).

⁵⁰ See generally Theresa W. Karle & Thomas Sager, *Are the Federal Sentencing Guidelines Meeting Congressional Goals?: An Empirical and Case Law Analysis*, 40 EMORY L.J. 393 (1991).

⁵¹ WRIGHTSMAN, *supra* note 49, at 373. Future study of judges' discretion in the sentencing process is warranted, given the recently revised sentencing guidelines. See *infra* notes 280-88 and accompanying text (discussing future research on sentencing).

⁵² WRIGHTSMAN, *supra* note 49, at 378-80 (noting impact of judges' age, experience, previous employment as district attorney, political party membership, and philosophy of punishment as predictors of magnitude of sentence imposed).

Ebbesen and Konecni observed more than 400 sentencing hearings to study empirically the factors that influence judges' sentencing decisions. Ebbe B. Ebbesen & Vladimir J. Konecni, *The Process of Sentencing Adult Felons: A Causal Analysis of Judicial Decisions* (providing empirical analysis of eight judges' sentencing behaviors and patterns), in *THE TRIAL PROCESS* 413, 431-58 (Bruce D. Sales ed., 1981). Similar to the approach of the model here, Ebbesen and Konecni sought to isolate the factors that accounted for the systematic variation among judges' sentences. Four factors accounted for the vast majority of

which judges' sentencing decisions are independent of legal or extralegal factors.⁵³

"G": competence and influence of participants variable

The model explores how the perceived competence or influence of different trial participants (e.g., judge or counsel) relates to other variables in the model, such as trial outcome or sentencing patterns.⁵⁴

II. THE STANFORD STUDY: TESTING THE MODEL IN JURY TRIALS

This Part highlights the findings from the Stanford Study of judges' behavior in criminal jury trials. Section A summarizes the Stanford findings from the videotaped analyses of judges' communicative behavior and discusses their relevance to the development of the on-line⁵⁵ coding of legal and extralegal measures in the Iowa Study. Section B highlights the simple relationships and Section C the cumulative relationships among the variables in the model explored in the Stanford Study.

A. Global and Micro Dimensions of Judges' Behavior

This Section describes the Stanford findings for judges' global styles and micro nonverbal behaviors in relating to their juries.⁵⁶ The findings provide the foundation for the on-line test of the model in the Iowa courts. Global style refers to the dimensions of judges' communicative and interpersonal behavior that are often conveyed independently of verbal content.⁵⁷ Although a particular global style may reflect a judge's orientation for relating to others during the trial, a

variation in sentencing, which may also be employed in subsequent tests of our model, including: (1) the type of crime; (2) the defendant's criminal history; (3) the status of the defendant between arrest and conviction (e.g., released on bail or held in jail); and (4) the probation officer's sentence recommendation. Of these factors, judges followed the recommended sentence of the probation officer in 84% of the cases studied. Ebbesen and Konecni concluded that the probation officer's recommendation is likely based on the other variables studied; that is, based on a knowledge of the defendants' criminal histories ("A" variable in our model) and on perceptions of the seriousness of the crime (potential "A" variable in our model). See also WRIGHTSMAN, *supra* note 53, at 382 (suggesting this is one reason why, in this study, judges' agreement with probation officers' recommendations were so high). Future tests of the model may also incorporate the sentencing recommendation of the probation officer as an additional post-verdict variable. See Shari Diamond, Exploring Sources of Sentence Disparity (providing empirical analysis of variables predicting judges' sentencing behaviors and patterns), in *THE TRIAL PROCESS*, *supra*, at 387.

⁵³ See Sue et al., *supra* note 46, at 352 (noting that sentencing judgments should not be related to jurors' degree of confidence in their verdicts); see also Hagan & Bumiller, *supra* note 6, at 10-18 (describing methodological difficulties in sentencing research, such as alternative measures of sentence severity and sampling problems).

⁵⁴ Peter D. **Blanck** & Robert Rosenthal, Mediation of Interpersonal Expectancy Effects: Counselor's Tone of Voice, 76 J. EDUC. PSYCHOL. 418 (1984). In these studies, less competent and less effective counselors were more prone to biasing effects, talking more negatively about others of perceived low potential and more positively (e.g., more warmly) about others of perceived high potential. More competent and interpersonally more effective individuals treated others of perceived high and low potential more equally. Similarly, in our work with psychotherapists, therapists' nonverbal behavior, as expressed only in tone of voice, is rated as significantly less anxious when they are rated as more competent by their supervisors. Peter D. **Blanck** et al., Talking to and About Patients: The Therapists Tone of Voice, in *NONVERBAL COMMUNICATION IN THE CLINICAL CONTEXT* 99, 129 (Peter D. **Blanck** et al. eds., 1986).

⁵⁵ The term "on-line" refers to the observing and recording of trial information by those present at the trial.

⁵⁶ Detailed statistical analyses and tables of the preliminary findings for the simple relationships of the variables in the model are described in *The Appearance of Justice*, *supra* note 1, at 119-136; *The Measure of the Judge*, *supra* note 1, at 672-74.

⁵⁷ See **Blanck** et al., *supra* note 54, 99-143 (providing overview of research program that analyzed five areas of interpersonal communication: descriptive, psychometric, interactional, competence, and trans-situational factors); Peter D. **Blanck** et al., Therapists' Tone of Voice: Descriptive, Psychometric, Interactional, and Competence Analyses, 4 J. SOC. & CLINICAL PSYCHOL. 154, 155-75 (1986) (summarizing research program on therapists' tone of voice); see also Figure 1, *supra* text accompanying note 40 (summarizing research model).

judge may show different global styles at different times depending on the circumstances of the trial process. For example, when responding to improper attorney behavior, a judge might show more directive or controlling behavior. Conversely, when dealing with a child witness, a judge might show more caring and patient behavior.

The Stanford Study is the first to delineate the dimensions of trial judges' global styles. The analyses yield four global dimensions of behavior: (1) "judicial," (2) "directive," (3) "confident," and (4) "warm."⁵⁸ These dimensions may be delineated further into those that appear more legally or procedurally oriented, as reflected by the judicial and directive dimensions, and those that appear more emotionally based, as reflected by the confident and warm global dimensions.⁵⁹

Analyses of the Stanford videotapes led to four preliminary conclusions about the "appearance" of judges' global dimensions of behavior:

- (1) A high rating on the judicial dimension indicates that the judge was perceived as professional, wise, competent, and honest. The judicial dimension is focused, perhaps in the broadest sense, on the appearance of judicial propriety and fairness.⁶⁰
- (2) A high rating on the directive dimension indicates that the judge was perceived as more dogmatic and dominant. The directive dimension typifies the qualities of the trial judge as a courtroom manager and as an administrator.⁶¹
- (3) A high rating on the confident dimension indicates that the judge was perceived as less anxious and less hostile. This dimension reflects the extent to which the judge appears emotionally comfortable, self-assured, and patient with others during the trial.⁶²
- (4) A high rating on the warm dimension indicates that the judge was perceived as more open-minded and empathic. This dimension reflects the extent to which the judge appears to be supportive of, and courteous toward, the trial participants.⁶³

The global dimensions help to predict other measures in the model; for instance, trial outcomes or judges' expectations. They also reflect interpretable and externally valid dimensions of judges' communicative styles that are consistent with prior case-oriented and clinically derived

⁵⁸ See *The Measure of the Judge*, supra note 1, at 657-62 (discussing principal components statistical methodology to analyze judges' behavior in more useful and practical ways by reducing number of variables to describe such behavior); see also EISENSTEIN & JACOB, supra note 46, at 24-28 (proposing four organizational goals that reflect the degree of adversarialness, judges' behavior, and patterns of case disposition: reducing conflict, avoiding uncertainty, processing cases, doing justice; these goals may relate to the four global styles, respectively: warm, confident, directive, and judicial).

⁵⁹ *The Measure of the Judge*, supra note 1, at 661-66.

⁶⁰ See Martin H. Redish & Lawrence C. Marshall, *Adjudicatory Independence and the Values of Procedural Due Process*, 95 *YALE L.J.* 455, 483-84 (1986) (describing core values of procedural due process to include "appearance" of an independent and fair adjudicator).

⁶¹ The "directive" dimension is similar to other case-study descriptions of so-called "managerial judging" techniques. See Judith Resnik, *Managerial Judges*, 96 *HARV. L. REV.* 376, 380, 445 (1982) (criticizing more active, managerial stance of judges and advocating a return to the classical judicial role); cf. Steven Flanders, *Blind Umpires-A Response to Professor Resnik*, 35 *HASTINGS L.J.* 505, 505-22 (1984) (criticizing Resnik's article for overstating the extent of judicial activity that is inconsistent with due process and for employing questionable approaches with established and acceptable practices and models in conducting her study). Future research should explore the relationship between judges' behavior and their methods and styles of case management.

⁶² *The Measure of the Judge*, supra note 1, at 662-65 (noting that a "confident" judge frames commands in the form of pleasant requests and is respectful of trial participants).

⁶³ The "warm" dimension is consistent with the style of positive regard for others advocated by the client-centered therapeutic school first advanced by Carl Rogers. See *The Measure of the Judge*, supra note 1, at 665-66 (describing analysis of therapeutic interaction and warm dimension).

descriptions of judges' behavior.⁶⁴

The Stanford Study also explored the micro nonverbal behaviors of trial judges. Examples of judges' micro behaviors include amount of eye contact with trial participants and frequency of smiles, hand movements, or head nods.⁶⁵ The micro behaviors assessed have been employed in studies of psychotherapy as well as in other studies of courtroom processes.⁶⁶

The Stanford Study demonstrates the predictive relationship between global styles and micro behaviors, highlighting the potential methodological contribution of the model.⁶⁷ The Stanford Study provides preliminary evidence that judges' micro behaviors alone predict with practical benefit those same judges' global styles.⁶⁸ The results show, for instance, that more engaged micro behaviors by the judge, such as eye contact and head nods directed at the jury, predict those same judges' judicial and directive global styles. Similarly, less engaged micro behaviors by the judge, such as less postural attention, predict those same judges' warm global style toward their juries.⁶⁹

The findings suggest methodological and economical shortcuts for researchers, courts, and practitioners interested in studying judges' behavior, attitudes, and decision making "on-line" during actual trials.⁷⁰ Even moderate relationships between the readily quantifiable micro behaviors and the more impressionistic global styles are of value to social scientists, legal researchers, courts, and practitioners.⁷¹ This is because the logistical and ethical problems often associated with videotaping and coding actual trials to assess judges' behavior are minimized.⁷²

⁶⁴ The National Conference of State Trial Judges describes the essential qualities of a good judge as including: graciousness, moral courage, reputation for fairness, mercy, patience, ability to communicate, decisiveness, innovation, open-mindedness, brevity, dignity, honesty, and integrity. THE JUDGE'S BOOK 31-38 (National Conference of State Trial Judges ed., 1989).

⁶⁵ See *infra* appendix B. Like the findings for the global dimensions, the Stanford Study provided a practical description of two basic constellations of judges' micro behavior. A first component of "engaged" micro behaviors emerged, with judges rated high on this component displaying more eye contact, more postural attention, and less self-touching. A second component of "emotional" micro behaviors emerged, with judges rated high on this component displaying more smiles, head nods, hand movements, and forward leans. The two micro constellations are conceptually similar to the "judicial" and "warm" global dimensions of judges' behavior.

⁶⁶ See generally *The Measure of the Judge*, *supra* note 1; see also Gordon D. Hemsley & Anthony N. Doob, *The Effect of Looking Behavior on Perceptions of a Communicator's Credibility*, 8 J. APPLIED SOC. PSYCHOL. 136 (1978) (rating witness less credible if he or she failed to maintain eye contact with lawyer); Bert Pryor & Charner Leone, *Behavioral Stereotypes of Deceptive Communication*, TRIAL, June 1981, at 14, 14-19, 70 (finding less eye contact, more backward leans, trunk swivel, leg movement, self-touching, gesturing, and speech errors associated with witnesses' attempts at deception).

For a review of the predictive value of micro behaviors in the courtroom context, see *The Measure of the Judge*, *supra* note 1, at 667-69.

⁶⁷ See *The Measure of the Judge*, *supra* note 1, at 674; see also *supra* note 57-64 and accompanying text (discussing global dimensions of judges' behavior).

⁶⁸ This suggestion echoes the large body of research demonstrating, for instance, the important effects of eye contact on social influence and credibility in courtroom testimony. *The Measure of the Judge*, *supra* note 1, at 674-75.

⁶⁹ *Id.* at 673-74.

⁷⁰ *Id.* This is true even where, as in the Stanford Study, analysis involves a relatively small sample of judges, all of whom knew that they were being videotaped and assessed by a naive group of observers. Moreover, the findings are particularly encouraging given the brief length of the judges' behavior that was rated. See Nalini Ambady & Robert Rosenthal, *Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis*, 111 PSYCHOL. BULL. 256, 256-58 (1992).

⁷¹ See *The Measure of the Judge*, *supra* note 1, at 676 (proposing ways in which the model and its results could be used in appellate review or as a training tool to modify judges' behavior); see also **Blanck** et al., *supra* note 54, at 127-29, 131-33 (detailing analogous line of study showing predictive value of micro behaviors in psychotherapeutic context).

⁷² See generally *Selecting Impartial Juries: Must Ignorance Be a Virtue in Our Search for Justice?* Keynote Address: The Impact of Television on the Jury System: Ancient Myths and Modern Realism, The Annenberg Washington Program Conference, May 11, 1990, 40 AM. U. L. REV. 623 (1991) (statement of Fred Graham) [hereinafter *Keynote Address*]; see also

Thus, researchers and court observers may use the more easily observed micro behaviors as an index of judges' global styles.⁷³

A. *Simple Relationships in the Model*

The model in the Stanford Study generated simple relationships that are summarized by the correlation between any two variables in the model. The model also generated cumulative relationships that are summarized by two or more variables predicting a third variable using multiple regression analyses. The Stanford findings are summarized next as a point for comparison with the findings of the Iowa Study that are set forth in Part IV below.⁷⁴

1. Background-Expectancy Relationships

This relationship describes how a judge's expectations for trial outcome may be predicted solely from the background variables of the trial participants. The Stanford Study suggests, for instance, that judges' beliefs about trial outcomes are related in predictable ways to defendants' criminal histories. Judges tend to expect a guilty verdict when defendants have more serious criminal histories and judges expect not guilty verdicts when defendants have less serious criminal histories.⁷⁵ Also, the Stanford Study shows a trend for judges to infer guilt when defendants are of a lower socioeconomic status.⁷⁶ These results suggest that although the background characteristics of defendants should have no legal bearing on guilt or innocence, they may influence judges' expectations for trial outcomes.

2. Background-Behavior Relationships

This relationship describes how defendants' background variables relate to judges' global and micro behaviors during trial. The Stanford Study shows that information about a defendant's criminal history, information that the jury is ordinarily not allowed to learn unless the defendant

INTERPERSONAL EXPECTATIONS, *supra* note 41, at 349-51 (discussing logistical and ethical problems associated with videotaping trials); **Blanck & Turner**, *supra* note 30, at 109-23 (stating that clinical field research is intervention that inevitably alters behavior and attitudes observed and recorded).

⁷³ See **Blanck & Turner**, *supra* note 30, at 113-15 (delineating micro behaviors studied and coding scheme attached to such behaviors). The long-term goal of the on-line coding system is to also explore the effectiveness of the model when employed by training or educational programs for judges devoted to the fine-grain videotape or on-line analysis of courtroom behavior. See generally *infra* part V; *The Measure of the Judge*, *supra* note 1, at 674.

⁷⁴ In the presentation of the preliminary findings, statistical significance is indexed by a probability that an observation would have been found if, in the population from which we sampled, the true correlation were zero. We typically present probability values (*p*) of .10 or less because these values are useful in assessing the types of variables under study here. See *The Appearance of Justice*, *supra* note 1, at 119-20 n.98.

In exploring the simple relationships of the model, we employ correlational analyses. The correlation coefficient (*r*) can take on values between -1.00 and +1.00. A value of -1.00 means that there is a perfect negative relationship, a value of +1.00 means there is a perfect positive relationship, and a value of .00 means that there is no linear relationship between the two variables in the model. Correlational analyses describe the predictive relationship between two variables but do not isolate the "causes" and "effects" of that relationship.

In exploring the complex relationships of the model, we employ multiple regression analyses. The Multiple R, or R², represents the relationship between a particular variable in the model (the criterion variable) and the set of predictor variables in the model. R² takes on values only between 0 and 1, with the former indicating no relationship and the latter indicating a perfect relationship between the variables. The F and t tests describe the level of confidence that the linear relationship between the criterion and predictor variables is not zero in the population. The df refers to the "degrees of freedom" required for statistical significance testing. See *The Measure of the Judge*, *supra* note 1, at 669.

⁷⁵ See *The Appearance of Justice*, *supra* note 1, at 120-21 (illustrating how defendants' prior records may affect judges' expectations for conviction or acquittal).

⁷⁶ *Id.* at 121.

takes the stand to testify, relates to judges' behavior when instructing their juries.⁷⁷ When delivering instructions for defendants with more serious criminal histories, judges' overt, verbal behavior is perceived as more judicial, directive, and warm,⁷⁸ but their nonverbal behavior is seen as relatively less judicial, directive, and warm.⁷⁹ The Stanford Study also shows that there is a tendency for raters to perceive judges' nonverbal behavior when delivering jury instructions as more professional and competent when defendants are of a higher socioeconomic status.⁸⁰ These results suggest that judges may sometimes "leak" or reveal to juries their underlying beliefs about defendants through nonverbal channels alone.⁸¹

3. Background-Outcome Relationships

Defendants' criminal histories tend to be related to trial outcomes: defendants with more serious criminal histories are more likely to be found guilty.⁸² Additionally, defendants of higher socioeconomic status received more guilty verdicts than defendants of lower socioeconomic status.⁸³ These results, viewed in combination with the tendency for the Stanford Study judges to nonverbally reveal criminal history, indicate that defendants' background variables might influence judges and juries more than many commentators previously assumed.⁸⁴

4. Background-Sentence Relationships

This relationship was not tested in the Stanford Study. It describes, however, the extent to which background variables predict the magnitude of the sentence imposed by the judge. The magnitude of the judge's sentence often will properly reflect the relevant criminal history of the defendant.⁸⁵ The analyses in the Iowa Study focus on other extralegal variables, such as the age, gender, race, and socioeconomic status of the defendant. There is no reason a priori to believe that extralegal background variables of the defendants or other trial participants should relate

⁷⁷ Id. at 121-24.

⁷⁸ See id. at 122 (describing the correlation between judges' communication styles and the seriousness of defendants' criminal histories).

⁷⁹ See id.

⁸⁰ Id. at 128.

⁸¹ See id. at 122-24 (detailing study results that support the theory that judges reveal personal opinions to juries through nonverbal communication).

⁸² See id. at 129 (providing statistical correlation between guilty verdicts and criminal histories); see also McElhaney, *supra* note 40 (noting that the most important decision for a defendant with prior convictions is whether to take the stand at trial).

⁸³ The Appearance of Justice, *supra* note 1, at 130 n.117. This was especially true for the first count against the defendant. Id.

⁸⁴ See id. at 130. Note also the background-judge/jury agreement ("A- E") relationship: This relationship describes how background variables may relate to the judge's views about the trial outcome-in the Stanford Study, about the jury verdict. Tests of this relationship suggest a slight trend for judges' disagreement with their juries' verdicts to be stronger when defendants had more serious criminal histories. Simple correlation for this relationship is $r = -.16$, not significant at $p < .10$. But this relationship may prove more apparent when assessed in the context of multiple regression analyses. See *infra* note 100 and accompanying text. We hypothesize, but have not yet tested, that in cases where the judge knows the defendant's prior criminal history and the jury does not, the judge will view a jury verdict of innocence to be overly lenient. This suggestion is consistent with Kalven and Zeisel's conclusion in *The American Jury* that in cases of judge/jury disagreement, juries tended to be viewed as more lenient than judges. See *infra* note 99.

⁸⁵ Insufficient data are available from the Stanford Study to test the relationship between defendants' criminal history and sentencing patterns. The Iowa Study also involved an insufficient number of defendants with prior criminal histories. See *infra* text accompanying note 112. Therefore, future study is needed to address this question, although archival analyses could be performed on existing data to address this issue.

For a review of research on sentencing patterns, see generally Ebbesen & Konecni, *supra* note 52 (suggesting that the relation between defendants' criminal histories and perceptions of the seriousness of the crime is predictive of judges' sentences). An interesting subset of cases to study further are those in which the defendant does not take the stand to testify. For this subset of cases, one might expect that defendants receive relatively lighter sentences. This finding should be particularly apparent in cases where the judge's initial expectations are for a not guilty verdict.

systematically to the magnitude of the sentence imposed by the judge. The existence of such a relationship would suggest potential bias and reversible error on the part of the trial judge.⁸⁶

5. Expectancy-Behavior Relationships

These relationships describe how judges' expectations for trial outcomes relate to their global and micro behaviors. The Stanford findings suggest that judges may reveal their beliefs to their juries through nonverbal channels alone. For instance, judges expecting a guilty verdict tend to be somewhat less judicial and warm in relating to their juries.⁸⁷ These results imply, but do not prove, that judges' verbal and nonverbal channels may convey messages concerning defendants' guilt or innocence.⁸⁸

6. Expectancy-Outcome Relationships

This relationship describes how judges' expectations for trial outcomes relate to actual trial outcomes. The Stanford findings suggest that judges' expectations for trial outcomes (as assessed by the judges themselves) do not accurately predict actual trial outcomes.⁸⁹ The Iowa Study revisits this issue by employing independent observers to make assessments, at different points during the trial, of judges' expectations for guilt or innocence.⁹⁰

7. Expectancy-Sentence Relationships

The Stanford Study did not test this relationship. The Iowa Study explores the extent to which the magnitude of the sentence reflects the expectations formed by the judge during the trial about the defendant's guilt or innocence.⁹¹ A finding may reflect impermissible bias when it shows a strong relationship between a judge's preconceptions for guilt during trial—an extralegal factor—and the magnitude of the sentence subsequently imposed by that same judge.⁹²

⁸⁶ For a discussion of reversible and harmless error, see *infra* notes 227-40 and accompanying text.

⁸⁷ See *The Appearance of Justice*, *supra* note 1, at 130-31.

⁸⁸ Intentionally or unintentionally, judges' beliefs may influence their verbal and nonverbal communication styles when relating to their juries, although on the written trial transcript or in response to a pencil-and-paper questionnaire, the judges may "appear" (or actually believe themselves to be) impartial. The jurors in the Stanford Study generally failed to report that the judges' nonverbal behavior conveyed the judges' expectations about what the trial outcomes should be. *Id.*; INTERPERSONAL EXPECTATIONS, *supra* note 41, at 342 (analyzing research process).

⁸⁹ See *The Appearance of Justice*, *supra* note 1, at 132-33. The simple correlation for this relationship is $r = -.20$, which is not significant at $p \leq .10$, but suggests a trend that expectations may be related to outcomes in predictable ways. See *infra* note 150 and accompanying text (discussing results of Iowa Study).

Nevertheless, the Stanford Study findings suggest that judges who send "expectancy effects" to juries should be held accountable, because biasing messages (at least as assessed by our methods) might not be an inevitable product of courtroom dynamics. INTERPERSONAL EXPECTATIONS, *supra* note 41, at 342. Empirical testing of the cumulative impact of the model (e.g., "A-B-C-E" predicting the "D" chain) is necessary to understand this relationship more completely.

⁹⁰ Note that the expectancy-judge/jury agreement ("B-E") relationships are not tested in the Iowa Study. See *supra* text accompanying notes 48-49. This relationship describes how judges' expectations prior to the trial outcome predict whether those same judges will agree or disagree with their juries' verdicts. Those judges expecting a guilty verdict prior to the trial outcome would probably be more likely to agree with a jury verdict of guilt than with a verdict of innocent. In the Stanford Study, the simple correlation for this relationship is $r = -.39$, $p < .05$. Judges who are convinced early in the trial that the defendant is guilty may likely reflect or "confirm" this attitude in the subsequent magnitude of their agreement with the jury's verdict. Moreover, the strength of the "B-E" relationship may be moderated by the severity of the defendant's criminal history, by the magnitude of the judge's expectations, and by whether or not the defendant testified. This suggestion would parallel Kalven and Zeisel's view in *The American Jury* that in cases where the judge has some knowledge of the defendant's criminal history that the jury does not have, the judge and jury may in fact be trying two different cases. KALVEN & ZEISEL, *supra* note 10, at 121. In other words, had the jury known what the judge knew, it would likely have agreed with the judge.

⁹¹ See generally Ebbesen & Konecni, *supra* note 52 (attempting to discern the decision-making strategies judges employ in matching sentencing options with individual offenders).

⁹² See *infra* notes 235-40 and accompanying text.

8. Behavior-Outcome Relationships

This relationship describes how judges' behavior alone may predict trial verdicts.⁹³ The Stanford results suggest a trend for judges' global styles to be less judicial and directive and their micro behaviors to be significantly more engaged when the jury reached a guilty verdict.⁹⁴ The Iowa Study again tests the extent to which judges' behavior predicts trial outcomes.⁹⁵ The Iowa Study also enables the study of behavior, evidence, and outcome in marginal and close cases.⁹⁶ The addition of the "C" variable thus enables an analysis of the strength of the evidence as a predictor of trial outcomes and of other trial process variables.⁹⁷ For example, in close cases in which the evidence toward guilt is relatively weak, judges' beliefs and behavior, defendants' background variables, or other extralegal factors likely play a relatively more important role in predicting trial outcomes.⁹⁸

9. Behavior-Sentence Relationships

This relationship describes the extent to which judges' global and micro behaviors at trial predict the imposition of more severe sentences. The Iowa Study provides a new avenue for study given the relative logistical ease with which judges' global and micro behaviors can be assessed.⁹⁹

A. Cumulative Relationships in the Model

⁹³ See *The Appearance of Justice*, supra note 1, at 133-36 (documenting study results showing correlation between judges' behavior and jury verdicts).

⁹⁴ See *id.* at 135. Further, simple correlation for the relationship between judges' engaged micro behavior (e.g., more eye contact) and trial outcome is $r = .28$, $p \leq .10$.

⁹⁵ See *infra* note 159-61 and accompanying text. Rosenthal summarized the results of the Stanford Study. "[I]f a judge believed a particular defendant to be guilty, sixty-four percent of the jurors found the defendant guilty. If the judge believed the defendant innocent, then only forty-nine percent of the jurors reported guilty. There was a fifteen difference (between sixty-four percent and forty-nine percent) that was, in a sense, in the mind and expectation of the judge." Panel One, supra note 3, at 1043 (statement of Robert Rosenthal). These results suggest that the judges may have conveyed their expectations through nonverbal channels to their juries.

⁹⁶ *The Appearance of Justice*, supra note 1, at 133-36. One encouraging general conclusion from the Stanford Study is that, in most cases where the judges' expectations for trial outcomes are conveyed to juries either verbally or nonverbally, jurors still tended to make their own independent assessments of the evidence. See *infra* notes 178-195 and accompanying text.

⁹⁷ See *infra* notes 181, 184 and accompanying text.

⁹⁸ See *infra* part IV.B.2 (presenting partial correlational analyses controlling for the strength of the evidence); cf. KALVEN & ZEISEL, supra note 10, at 134-35 (analyzing evidence in terms of "close" and "clear" cases).

Note also the behavior-judge/jury agreement ("C-E") relationship, which describes how a judge's global styles and micro behaviors at trial may predict (or be predicted by) that judge's views about the trial outcome. Preliminary tests of this relationship in the Stanford Study suggest a trend indicating that judges' engaged micro behaviors are related to their views about trial outcome. Although the simple correlation between the "C" and "E" variables was not significant, the partial correlations derived from exploratory multiple regression analyses suggested a strong relationship: partial $r = .26$, $p < .04$. This relationship supports the suggestion that judges' behavior alone may reflect their views about the perceived "correctness" of their jury's ultimate conclusion.

⁹⁹ The outcome-judge/jury agreement ("D-E") relationship explored in the Stanford Study showed that trial outcome is predicted by a knowledge of the magnitude of judges' agreement/disagreement with their juries' verdicts. Simple correlation for this relationship is $r = .63$, $p < .01$. Judges are more likely to agree with guilty verdicts and to disagree with not guilty verdicts. On its face, this result supports Kalven and Zeisel's general conclusion that judges tend to view jury results as more lenient than their own. Further research is being conducted to isolate this result, given that in our research (unlike that of *The American Jury*) the judges completed the questionnaires both before and after they had knowledge of the jury's verdict. Kalven and Zeisel recognized that a possible source of methodological bias existed in their study because they could not be sure that their participating judges completed the questionnaire responses before actually hearing their jury's verdict. See KALVEN & ZEISEL, supra note 10, at 52. It will be interesting to explore the extent to which our results parallel those of Kalven and Zeisel's, given the different methods. See *The Appearance of Justice*, supra note 1, at 157-58.

The model is most powerful and predictive when examining combinations of variables. Multiple regression analyses are used to explore cumulative effects in the model. Regression analyses enable detailed assessments of the relationship between a set of independent predictor variables with one criterion (i.e., dependent) variable in the model.¹⁰⁰

The Stanford Study employed several variables in the model as predictors of trial outcome. One combination tested the extent to which trial outcomes are predicted by the set of variables, including defendants' criminal histories, judges' expectations for trial outcome, engaged micro behaviors of the judges, and magnitude of judge/jury agreement/disagreement as to verdict. Initial results suggest that judges' expectations, judges' engaged micro behavior, and judge/jury agreement together predict trial outcomes better than any single variable in the model alone.¹⁰¹

Similarly, the Iowa Study uses regression analyses to explore the "A- B-C-C" chain to predict trial outcome and sentencing. The analyses illustrate how study of other combinations of variables shed light on the long-standing observation that legal and extralegal factors influence trial outcomes and sentencing. Additionally, despite the growing tendency to videotape and televise actual trials,¹⁰² the Iowa test of the model illustrates how researchers, courts, and practitioners may systematically calibrate on-line courtroom behavior. Thus, the model may be useful for assessing courtroom behavior through videotape or on-line analyses tested in the Stanford and Iowa Studies.¹⁰³ As Part V suggests, videotape and on-line analyses may become standard tools in the calibration of the permissible limits of judges' behavior and of trial error. Such calibrations would not be possible through analysis of the written transcript or trial record alone.¹⁰⁴

III. THE IOWA STUDY: TESTING THE MODEL IN BENCH TRIALS

The Iowa Study explores: (1) the legal and extralegal factors that influence decision making in bench trials; (2) the import to the model of the strength of the evidence variable; (3) judges' behavior over the course of entire bench trials, rated at several points and involving interactions with different trial participants;¹⁰⁵ and (4) the validity and practical value of on-line coding for

¹⁰⁰ For a review of multiple regression techniques, see JACOB COHEN & PATRICIA COHEN, APPLIED MULTIPLE REGRESSION/CORRELATIONAL ANALYSES FOR THE BEHAVIORAL SCIENCES 7 (2d ed. 1983) (explaining that multiple regression analyses describe the relationships between a complex set of predictor variables and a single criterion variable). The term "cumulative chain," as used herein, is meant to describe the additive relationship of the variables and is not meant to have statistical relevance.

¹⁰¹ R^2 for this complex relationship = .7420, $F = 20.10$, $df(3, 30)$, $p < .0001$. For this equation, partial r for "B" variable = .67, $p < .001$; for "C" variable = .39, $p < .05$; and for "E" variable = .64, $p < .001$. Partial r for "A" variable in this chain = .25, $p = .17$.

¹⁰² See generally Keynote Address, *supra* note 72 (describing efforts to make trial proceedings available through television).

¹⁰³ But see Thomas Grisso et al., Standards in Research, 46 AM. PSYCHOLOGIST 758, 762 (1991) (providing reasons why research psychologists should proceed with extreme caution when employing videotaped data); Peter D. Blanck, The "Process" of Field Research in the Courtroom: A Descriptive Analysis, 11 LAW & HUM. BEHAV. 337, 349-51 (1987) (discussing due process problems and concerns of disruption to the trial process when videotaping trials for research purposes); Peter D. Blanck et al., Scientific Rewards and Conflicts of Ethical Choices in Human Subjects Research, 47 AM. PSYCHOLOGIST 959 (1992) (viewing research ethics not as an affront to the integrity of sound research but as presenting an opportunity for scientific rewards).

¹⁰⁴ The present study also furthers the analysis of trial processes by testing the psychometric value (reliability, consistency, and validity) of on-line coding methods. See *infra* appendix C. Another step in the test of the model will be to validate the degree to which videotape and on-line analyses predict trial outcome and other measures in the same trials.

¹⁰⁵ Cf. Panel One, *supra* note 3, at 1045-47 (statement of Judge LaDoris Cordell) (discussing the importance of studying trial judges' behavior when interacting with counsel and witnesses).

assessing trial outcomes and sentencing.¹⁰⁶

A. Research Strategy and Design: Trials and Participants

Trials. The study was conducted in the Johnson County Magistrate Court in Iowa. A magistrate presided over each of fifty-two trials that were studied during a twenty-week period.¹⁰⁷ Three magistrates, two men and one woman, participated in the study. The three magistrates were studied in twenty-one, seventeen, and fourteen trials, respectively, for a total of fifty-two trials.

The charges in the trials included public intoxication, failure to maintain control of a vehicle, speeding, assault, and theft.¹⁰⁸ Of the fifty-two trials studied, forty-one resulted in guilty verdicts (79%) and eleven in not guilty verdicts (21%). For those defendants receiving guilty verdicts, twelve defendants (29%) were awarded the maximum sentence or scheduled penalty.¹⁰⁹ The typical maximum sentence for the first charge in the trials was thirty days in jail or a \$100 fine.¹¹⁰

Defendants. Seventy-nine percent of the defendants were men. The average age of the defendants was twenty-seven years, with a range of twenty to fifty years. Defendants' socioeconomic status ranged from low (14%) to medium (81%) to high (6%).¹¹¹ Only two cases involved defendants with prior criminal histories; therefore the Iowa Study did not test this background variable.

Lawyers. Counsel for the state either worked for the district attorney's office or were legal interns. About half of the prosecution counsel were men (44%), and half were women (56%). Most defense counsel were men (88%). The majority of defendants, however, represented themselves.¹¹²

¹⁰⁶ See *infra* appendix C; see also *The Appearance of Justice*, *supra* note 1, at 113-14 (discussing internal and external validity and practical value of courtroom research). The data for the Iowa Study were collected in conjunction with Christopher Traynor, *A Study of Magistrates in Bench Trials* (1991) (unpublished undergraduate honors thesis, University of Iowa). The magistrate court proved to be a cost-effective site for the test of the extended model in bench trials. The trials studied lasted approximately one hour and, typically, four trials were scheduled each day.

¹⁰⁷ Data is presented and tested for all trials except where it is not available, in which analyses the sample size, or "n," is presented and tested as less than 52. Magistrates are referred to throughout this Article as trial judges or judges.

¹⁰⁸ The breakdown of the charges are: 33% (17) failure to maintain control of vehicle, 31% (16) speeding, 6% (3) assault, 10% (5) theft, and 20% (10) other (such as public intoxication and criminal trespass). See *infra* note 195 (discussing need for future analyses on seriousness of offense).

¹⁰⁹ The sentence variable ("F") is calculated by dividing the sentence received by the maximum possible sentence for that charge. See *supra* notes 49- 53 and accompanying text. Other findings for magnitude of sentence are: 5% received half the maximum sentence, 10% received additional penalties beyond the maximum sentence, and the remaining defendants received lesser sentences. The analyses were conducted only for the first major charge against the defendant. See *infra* appendix B (trial data collection measures).

¹¹⁰ In many cases, the "scheduled fine" is typically the imposition of the \$100 penalty.

¹¹¹ Cf. EISENSTEIN & JACOB, *supra* note 46, at 181-82 (discussing measurement problems with independent variable of socioeconomic status). Police reports are often inaccurate measurements of socioeconomic status. Another measure of socioeconomic status might be whether the defendant has retained counsel. *Id.* at 182.

¹¹² Eighty-five percent (44 of 52) of the defendants represented themselves. Future analyses of the model will address in greater detail the impact of a pro se defense on trial processes and judge decision making. Preliminary analyses were conducted separately, however, for trials in which defendants were pro se and represented by counsel. Results show a higher proportion of guilty verdicts for cases in which defendants were pro se (Chi Square Test = 4.717, $p = .03$). Also, judges were rated as expecting more guilty verdicts for pro se defendants relative to those represented by counsel ($t = 3.26$, $p = .002$), and the strength of evidence was rated as stronger toward guilt in trials involving pro se defendants ($t = -2.74$, $p = .008$). Consistent with the

B. Research Strategy and Design: On-Line Ratings of the Trials

Two independent observers conducted the on-line coding of each trial.¹¹³ Each observer watched and coded an entire trial, with coordination between the observers as to the timing of the ratings. The first five minutes of several critical segments of the trials (e.g., witness testimony, closing arguments, and deliberations) were coded independently.¹¹⁴

The on-line rating method enabled the collection of information on the following variables:

- a. Background Variables: age, gender, race, and socioeconomic status of the defendant, judge, prosecuting attorney, and defense attorney.
- b. Expectancy Variables: ratings of the judges' expectations for trial outcome made at three points during the trial: (1) witness testimony (direct and cross-examination), (2) closing arguments, and (3) the judges' deliberation periods before sentencing.
- c. Communicative Variables: ratings of the judges' seven micro behaviors and four global styles during: (1) witness testimony,¹¹⁵ (2) closing arguments, and (3) the judges' deliberation periods.
- d. Strength/Complexity of Evidence Variables: ratings at the conclusion of the trial.¹¹⁶ Observers assessed the strength of the evidence that would lead a fact finder to reach a decision of guilty or not guilty.¹¹⁷ Observers assessed the complexity of the evidence in terms of the legal technicalities, sophisticated analyses of physical evidence, or specialized knowledge or opinions of professionals (e.g., physicians or psychologists).¹¹⁸
- e. Outcome Variables: the finding of guilty or not guilty.

general trends reported below, the magnitude of the sentence imposed did not relate to pro se cases versus those in which defendants were represented by counsel ($t = -.128, p = .90$). The regression models below exclude defense counsel variables because of their small sample size.

¹¹³ Ten individuals, six men and four women, rated the 52 trials. The observers were paired randomly to view a particular trial. A summarized version of the data collection instrument is provided in Appendix B.

¹¹⁴ See *infra* appendix B. The coding of the first five minutes for a segment of the trial reflects the trend in psychotherapeutic research to sample approximately one to five minutes of the therapist's speech. The Appearance of Justice, *supra* note 1, at 118 n.97 (citing studies); ROBERT E. PITTENGER ET AL., *THE FIRST FIVE MINUTES* (1960) (demonstrating the verbal and nonverbal richness of the first five minutes of the therapeutic interview); see also Ambady & Rosenthal, *supra* note 70, at 252-70 (showing high accuracy of prediction of social outcomes based on short observations of verbal and nonverbal expressive behavior); Panel One, *supra* note 3, at 1043 (statement of Robert Rosenthal) (discussing Ambady & Rosenthal study). The Iowa Study enables an assessment of the psychometric factors to be considered in "live" courtroom research. There are many unanswered questions, however, about the reliability, validity, and consistency of segments of observations of the trial process. The analyses aid in assessments of how much of judges' behavior, or of other aspects of the trials, must be studied to generalize with confidence to other behaviors or segments in other trials. The Iowa Study also enables an analysis of the consistency of judges' behavior across the trial segments; for example, during witness testimony, during closing arguments, and during the deliberation period. Limited data was available for opening statements, so this segment was not included in the analyses. Finally, the study enables analyses of trial segments during which judges may be most likely to convey unintended or intended influence. This line of study may eventually enable the identification of critical points in the trial during which the behavior of the judge, versus the presentation of the evidence, has a significant impact on trial outcome or sentencing. See *infra* notes 225-40 and accompanying text (discussing appellate court assessment of trial error).

¹¹⁵ The judges' rulings on objections (e.g., sustained, overruled) were also noted during witness testimony. See *infra* appendix B. The results of these observations are not presented herein.

¹¹⁶ The evidentiary ratings were made after the judge determined guilt or innocence. Future study will need to assess the strength of the evidence during the trial to isolate the effect of trial outcome from ratings of the evidence.

¹¹⁷ The primary focus of the present analyses is on the strength of the evidence variable. Subsequent analyses are being conducted related to the complexity and quality of the evidence.

¹¹⁸ Eight of the ten observers were law students, and the other two were senior undergraduate pre-law majors. All observers were familiarized with the implications of various types of evidence.

- f. Sentence Imposed Variables: the magnitude of the sentence imposed.¹¹⁹
- g. Competency/Influence of the Trial Participants Variables: ratings at the end of the trial of judges', counsels', and witnesses' competence and influence.

IV. THE IOWA STUDY: FINDINGS AND IMPLICATIONS

This Part sets forth the findings for the test of the model in Iowa bench trials.¹²⁰ Section A summarizes the results for the simple relationships in the model. Section B sets forth in greater detail the findings for the cumulative effects of the legal and extralegal variables.

A. Simple Relationships in the Model

The simple bivariate relationships in the model are explored in this Section. These relationships are summarized in Figure 1 above.

1. Background-Expectancy Relationships

This relationship describes how judges' expectations for trial outcomes relate to defendants' extralegal background variables. The findings show that judges' expectations for guilt are stronger for older defendants, while judges' expectations for a finding of not guilty tend to be more prevalent for younger defendants.¹²¹ This finding suggests the usefulness of including the background variable of defendant age in subsequent cumulative tests of the model. No other defendant background variables related to judges' expectations for trial outcomes.¹²²

2. Background-Behavior Relationships

This relationship describes how defendant background variables relate to judges' global and micro behaviors during trial. Several findings are of interest. First, the defendant's age relates to the judge's behavior: when defendants are older, judges tend to display less forward leans,¹²³ yet appear warmer¹²⁴ during trial. Second, the defendant's gender relates to the judge's behavior: judges tend to display more self-touching behaviors when defendants are male.¹²⁵ Interestingly, during trials with male defendants, judges tend to appear more judicial,¹²⁶ directive,¹²⁷ warm,¹²⁸ and confident.¹²⁹

The findings are consistent with the Stanford results suggesting that judges may sometimes "leak" or reveal to trial participants-defendants and counsel in the present study-their underlying

¹¹⁹ The magnitude was computed by dividing the actual sentence by the potential maximum sentence.

¹²⁰ The analyses of the simple and cumulative relationships in the model are averaged over the course of entire trials. The results of the psychometric analyses set forth in Appendix C-showing relatively high observer reliability and consistency-enable the analyses to be averaged over trial segments. This, in turn, allows for more readily interpretable and practical assessments of the findings.

¹²¹ For this relationship, $r = -.23$, $p = .11$, $n = 52$.

¹²² Analyses are on file with the author.

¹²³ $r = -.21$, $p = .13$, $n = 52$.

¹²⁴ $r = .24$, $p = .08$, $n = 52$.

¹²⁵ $r = -.24$, $p = .09$, $n = 52$.

¹²⁶ $r = -.22$, $p = .11$, $n = 52$.

¹²⁷ $r = -.24$, $p = .09$, $n = 52$.

¹²⁸ $r = -.19$, $p = .18$, $n = 52$ (suggesting a trend).

¹²⁹ $r = -.20$, $p = .16$, $n = 52$ (suggesting a trend). The small sample size of men ($n = 2$) and women ($n = 1$) judges does not allow for meaningful analysis of the relation of judge gender and defendant gender with communication style of the judge.

beliefs or expectations about aspects of the trial through nonverbal channels alone.¹³⁰ The cumulative analyses shed additional light on the extent to which these judges may leak, through their global and micro behaviors, expectations for trial outcomes and the effect this may have on trial outcomes and sentencing.

3. Background-Evidence Relationships

These findings show that the evidence tends to support guilty verdicts more frequently in trials of older defendants than in trials of younger defendants,¹³¹ and in trials of female defendants than in trials of male defendants.¹³² It is possible that the results for defendant age or gender reflect societal biases or the gender role socialization of the observers themselves.¹³³ Further study of the relation between the perceived and actual strength of the evidence and trial participants' background variables is warranted. The present findings imply that, although several of the background characteristics of defendants should have no legal bearing on the fact-finder's perception of the evidence or of ultimate guilt, background variables may relate in yet unidentified ways to the appearance of justice in the courtroom.¹³⁴

4. Background-Outcome Relationships

This relationship shows that older defendants are more likely to receive guilty verdicts, and younger defendants are more likely to receive not-guilty verdicts.¹³⁵ The relative impact of defendant age and gender, as central background variables in the model, is discussed further in the multivariate analyses.

5. Background-Sentence Relationships

This relationship describes how defendant background variables relate to the magnitude of the sentence imposed by the judge. Importantly, no significant relationships are found between defendant age or gender and the magnitude of the sentence imposed.¹³⁶ This nonfinding suggests that the sentencing patterns of these judges are not predicted solely by defendants' extralegal factors.

6. Background-Competence/Influence Relationships

In trials of older, as compared to younger, defendants, prosecution counsel are rated as more competent.¹³⁷ Also, in trials of older defendants, defense counsel tend to be rated as less

¹³⁰ See *The Appearance of Justice*, supra note 1, at 122-24 (detailing study results that support theory that judges reveal personal opinions to juries through nonverbal communication); see also Allen J. Hart, *On the Sobriety of Judges: Nonverbal Influence in the Courtroom* (1991) (unpublished Ph.D. dissertation, Harvard University) (finding that in mock jury trial scenarios, biased judicial behavior influenced jurors to return verdicts in same direction as judges' bias and to be more confident in their verdicts when they agreed with the judges' bias).

¹³¹ $r = -.34$, $p = .02$, $n = 52$.

¹³² $r = -.22$, $p = .12$, $n = 52$. No "A-B" relationships occurred involving ratings of the complexity of the evidence.

¹³³ See *The Appearance of Justice*, supra note 1, at 127-28 (discussing findings for gender of trial judges in Stanford Study).

¹³⁴ See *The Appearance of Justice*, supra note 1, at 119-21 (discussing "A-B" relationship in Stanford Study); see also infra part V.B & C (discussing research on trial error and sentencing).

¹³⁵ $r = .36$, $p = .01$, $n = 52$. Although defendant gender was not related significantly to trial outcome ($r = .15$, $p = .28$, $n = 52$), the trend shows that women defendants are somewhat more likely to receive a guilty verdict.

¹³⁶ For age, $r = -.06$, $p = .72$, $n = 40$. For gender, $r = -.19$, $p = .24$, $n = 40$ (showing a slight trend that men received greater magnitude of sentence).

¹³⁷ $r = .23$, $p = .10$, $n = 52$. This result is consistent with the pattern of results that older defendants are more likely to be found guilty. See supra notes 131-35 and accompanying text.

influential.¹³⁸ No such relationships emerge with defendant gender. However, in trials involving defendants of lower socioeconomic status, judges are rated as more competent and influential.¹³⁹ These findings suggest that defendant age and socioeconomic status might sometimes relate to perceptions of judges' or counsels' competence or effectiveness.¹⁴⁰

7. Expectancy-Behavior Relationships

This relationship describes how judges' expectations for trial outcomes relate to their global and micro behaviors during trial. The findings explore the suggestion that judges sometimes reveal their beliefs to trial participants through nonverbal channels alone.¹⁴¹ Consistent with the Stanford results, when judges expect guilty verdicts, they tend to show fewer smiles and postural changes during trial.¹⁴² The converse is true when judges expect not-guilty verdicts. No "B-C" relationships emerge, however, with the four global styles.¹⁴³

8. Expectancy-Evidentiary Relationships

This tests the relation between judges' expectations and the strength and complexity of the evidence at trial. Strong predicted relationships emerge. First, when judges expect guilty verdicts, the strength of the evidence is rated as strong toward guilt.¹⁴⁴ Conversely, when judges expect not-guilty verdicts, the evidence is rated as strong toward innocence.

The finding that judges' expectations for trial outcomes relate to the strength and direction of the evidence suggests that the judges and independent observers—who may be analogous to shadow jurors—in this study agree substantially as to the impact of the evidence.¹⁴⁵ It is not clear, however, whether judges' expectations influence the perceived strength of the evidence or vice versa. Judges' expectations formed during trials about ultimate trial outcomes are likely influenced by many factors. The multivariate analyses address this issue.

Also, when judges expect guilty verdicts, the evidence is rated as less complex.¹⁴⁶ Conversely, when judges expect not-guilty verdicts, the evidence is rated as more complex. The self-selecting nature of the cases that go to trial, versus those that are plea bargained or settled, could partially account for this result. That is, the cases that are tried are, in fact, ones in which the defendant is more likely to be guilty (i.e., the case is rated as less complex). Judges, knowing this, may expect trials to be less complex when the evidence is strong toward guilt. This suggestion is consistent with the Stanford findings where judges tended to expect guilty verdicts when defendants had relatively more severe criminal histories.¹⁴⁷ It is also consistent with Kalven and Zeisel's observation in *The American Jury* that judges are sometimes unable to avoid being influenced by

¹³⁸ $r = -.60$, $p = .12$, $n = 8$ (Note that this result is based on a small sample size.).

¹³⁹ $r = -.29$, $p = .04$, $n = 52$; $r = -.21$, $p = .13$, $n = 52$, respectively.

¹⁴⁰ Future uses of the model will be directed toward empirical assessments of "effective" assistance of counsel.

¹⁴¹ *The Appearance of Justice*, supra note 1, at 130-31.

¹⁴² For smiles, $r = .24$, $p = .10$, $n = 52$. For postural changes, $r = .21$, $p = .15$, $n = 52$ (suggesting only a trend in this direction).

¹⁴³ For smiles, $r = .24$, $p = .10$, $n = 52$. For postural changes, $r = .21$, $p = .15$, $n = 52$ (suggesting only a trend in this direction).

¹⁴⁴ $r = .57$, $p = .0001$, $n = 52$.

¹⁴⁵ Cf. *Empirical Research*, supra note 1, at 798-802 (discussing judge/jury agreement/disagreement findings and relation to *The American Jury* results).

¹⁴⁶ $r = .44$, $p = .0011$, $n = 52$.

¹⁴⁷ *The Appearance of Justice*, supra note 1, at 120. These results also could be affected by the types of cases that were studied (e.g., misdemeanors versus felonies).

defendant background variables.¹⁴⁸

9. Expectancy-Outcome Relationships

This relationship describes how judges' expectations for trial outcomes relate to actual trial outcomes.¹⁴⁹ Not surprisingly, the findings show that judges' expectations for guilt predict actual trial outcomes of guilt.¹⁵⁰ The primary issue explored in the Stanford Study was the extent to which judges' expectations are conveyed to jurors in impermissible ways. The findings for the Iowa Study of bench trials imply that judges' expectations formed during trials may be conveyed not only to independent courtroom observers, but to other trial participants as well, such as counsel and witnesses.

10. Expectancy-Sentence Relationships

Importantly, the magnitude of the sentence imposed by these judges does not reflect their expectations, formed during trial, about the defendants' guilt.¹⁵¹ This finding suggests that any measurable preconceived notions developed by these judges during bench trials about the defendants' guilt did not relate to their subsequent sentencing patterns.¹⁵²

11. Expectancy-Competence/Influence Relationships

This explores the relation between judges' expectations for trial outcomes and the perceived competence and influence of the judge and counsel. When judges expect a guilty verdict, they are rated as less influential.¹⁵³ Also, when judges expect a guilty verdict, prosecution counsel are rated as more competent.¹⁵⁴ More study is required to understand the complex relationship between competence of counsel and judges' expectations for trial outcomes.¹⁵⁵

12. Behavior-Evidentiary Relationships

This shows the relation among the micro and global behaviors and the strength of the evidence. When the evidence is rated as strong toward guilt, judges show less eye contact and fewer smiles¹⁵⁶ yet are rated as more judicial, directive, and warm toward trial participants.¹⁵⁷ The findings suggest that strong evidence of guilt may relate to displays of positive (e.g., fair) global

¹⁴⁸ See KALVEN & ZEISEL, supra note 10, at 124-27. "In addition to [the judge's] wide experience with the likelihood that the defendant before him is guilty, the judge is exposed to prejudicial information which the law, in its regard for the right of the defendant, aims to screen out of the evaluation of his guilt or innocence. The law's ideal in these situations may be something of a libertarian luxury." Id. at 127.

¹⁴⁹ The Appearance of Justice, supra note 1, at 132-33 (noting that the simple correlation for this relationship is $r = -.20$, which is not significant at $p \leq .10$, but suggests a trend that expectations may be related to outcomes in predictable ways).

¹⁵⁰ $r = -.46$, $p = .0006$, $n = 52$. Part V addresses when judges' expectations, which, in fact, impermissibly affect the course and outcome of trials, might constitute reversible error. See infra notes 245-62 and accompanying text (discussing trial and structural errors).

¹⁵¹ $r = -.16$, $p = .32$, $n = 40$.

¹⁵² Further research is needed, however, as almost two-thirds of the defendants in these trials did not receive scheduled penalties.

¹⁵³ $r = .36$, $p = .009$, $n = 52$. For judge competence, $r = -.14$, $p = .33$, $n = 52$ (nonsignificant result). This finding is consistent with the view that when the evidence is strong toward guilt the trial is less complex.

¹⁵⁴ $r = -.37$, $p = .007$, $n = 52$. Also, when judges expected guilty verdicts, defense counsel were rated as less influential ($r = .55$, $p = .16$, $n = 8$).

¹⁵⁵ One methodological shortcoming of the analysis here is that the observers' perceptions of the trial participants' competence and influence may be affected by their knowledge of the actual verdict, since these ratings are made at the close of the trial.

¹⁵⁶ For eye contact, $r = .31$, $p = .03$, $n = 52$; for smiles, $r = .22$, $p = .12$, $n = 52$. Also, when the evidence is rated as more complex, judges show more postural changes: $r = .24$, $p = .09$, $n = 52$.

¹⁵⁷ For judicial, $r = -.20$, $p = .16$, $n = 52$ (suggesting a trend); for directive, $r = -.23$, $p = .11$, $n = 52$; for warm, $r = -.31$, $p = .02$, $n = 52$.

styles by judges at trial.¹⁵⁸

13. Behavior-Outcome Relationships

This relationship describes how judges' behavior predicts their verdicts in bench trials.¹⁵⁹ Consistent with the trends in the Stanford Study, when the verdict is guilty, judges show less eye contact and fewer smiles and postural changes (i.e., they are more negative), yet are globally warmer in relating to trial participants (i.e., they are attempting to appear neutral).¹⁶⁰ These findings, which are explored further in the multivariate analyses, indicate identifiable relationships between judges' behavior during bench trials and their subsequent guilty verdicts.¹⁶¹

14. Behavior-Sentence Relationships

This relationship describes how judges' behavior relates to the magnitude of the sentence.¹⁶² The findings show that when the magnitude of the sentence is greater, judges show a trend for fewer postural changes, but are rated as warmer (again, reflecting attempts to appear neutral).¹⁶³

15. Behavior-Competence/Influence Relationships

This shows the relation among the micro and global behaviors and the perceived competence and influence of the trial participants. The findings may be summarized as follows: First, when judges are perceived as more competent (e.g., engaged), they display more head nods, hand movements, postural changes, and self-touching and are rated as more judicial, directive, warm, and confident.¹⁶⁴ When judges are perceived as more influential, they display more smiles, forward leans, and postural changes and are rated as more judicial and confident.¹⁶⁵ Second, when prosecution counsel are perceived as more competent and influential, judges display more head nods, hand movements, and self-touching and are rated as more judicial, directive, warm, and confident.¹⁶⁶ Third, when defense counsel are perceived as more competent, judges display

¹⁵⁸ It will be important to reexamine this relationship in jury trials, where the judge may reflect to jurors (through nonverbal channels alone) his or her views about the evidence presented at trial. In turn, this dynamic may impact counsels' trial strategies or the jury's decision-making process.

¹⁵⁹ See *The Appearance of Justice*, supra note 1, at 133-34 (documenting study results showing correlation between judges' behavior and jury verdicts).

¹⁶⁰ $r = -.26$, $p = .06$, $n = 52$ (for eye contact); $r = -.31$, $p = .03$, $n = 52$ (for smiles); $r = -.21$, $p = .13$, $n = 52$ (for postural changes); $r = .30$, $p = .03$, $n = 52$ (for warm).

¹⁶¹ Empirical Research, supra note 1, at 796 (noting that the simple correlation for the relationship between judges' engaged micro behavior (e.g., between more eye contact and trial outcome) is $r = .28$, $p \leq .10$).

¹⁶² Cf. *The Appearance of Justice*, supra note 1, at 133-34 (documenting study results showing correlation between judges' behavior and jury verdicts).

¹⁶³ $r = -.23$, $p = .16$, $n = 40$ (for postural changes); $r = .33$, $p = .04$, $n = 40$ (for warm). Judges' global and micro behaviors at trial, reflecting relatively less warm or more directive and engaged attitudes, were hypothesized to relate to the imposition of more severe sentences. The simple results do not support this prediction.

¹⁶⁴ $r = .43$, $p = .0015$, $n = 52$ (for nods); $r = .35$, $p = .012$, $n = 52$ (for hand movements); $r = .23$, $p = .10$, $n = 52$ (for postural changes); $r = .30$, $p = .04$, $n = 52$ (for self-touching); $r = .75$, $p = .0001$, $n = 52$ (for judicial); $r = .24$, $p = .08$, $n = 52$ (for directive); $r = .60$, $p = .0001$, $n = 52$ (for warm); $r = .75$, $p = .0001$, $n = 52$ (for confident).

¹⁶⁵ $r = .40$, $p = .0035$, $n = 52$ (for smiles); $r = .24$, $p = .09$, $n = 52$ (for forward leans); $r = .42$, $p = .002$, $n = 52$ (for postural changes); $r = .30$, $p = .04$, $n = 52$ (for judicial); $r = .34$, $p = .02$, $n = 52$ (for confident).

¹⁶⁶ For competent: $r = .28$, $p = .05$, $n = 52$ (for nods); $r = .20$, $p = .16$, $n = 52$ (for hand movements); $r = .25$, $p = .07$, $n = 52$ (for self-touching); $r = .56$, $p = .0001$, $n = 52$ (for judicial); $r = .42$, $p = .002$, $n = 52$ (for directive); $r = .51$, $p = .0001$, $n = 52$ (for warm); $r = .47$, $p = .0004$, $n = 52$ (for confident).

For influential: $r = .37$, $p = .0075$, $n = 52$ (for nods); $r = .31$, $p = .03$, $n = 52$ (for hand movements); $r = .39$, $p = .005$, $n = 52$ (for self-touching); $r = .53$, $p = .0001$, $n = 52$ (for judicial); $r = .24$, $p = .10$, $n = 52$ (for directive); $r = .46$, $p = .0007$, $n = 52$ (for warm); $r = .57$, $p = .0001$, $n = 52$ (for confident).

more hand movements and are rated as more judicial, directive, and confident.¹⁶⁷ Defense counsel who are perceived as more influential also receive more smiles from the judge.¹⁶⁸ Taken together, these findings imply that the intensity of courtroom interactions among trial participants may, in part, be mediated by perceptions of the participants' competence and influence. Thus, judges' perceptions of the competence of counsel may be reflected in their behavior during the trial (for example, in their reactions to counsels' objections or to witness testimony).¹⁶⁹

16. Outcome-Competence/Influence Relationships

This is the relation of trial outcomes to the competence and influence ratings of the trial participants. The findings show that when the verdict is guilty, judges are rated as less influential, prosecution counsel are rated as more competent, and defense counsel are rated as less influential.¹⁷⁰ This pattern of results suggests, predictably, that successful counsel are perceived as more competent.¹⁷¹

17. Sentence-Competence/Influence Relationships

This shows the relation between the magnitude of the sentence and ratings of the trial participants. Consistent with prior findings, when the magnitude of the sentence is greater, judges tend to be rated as less influential.¹⁷² The strength of the relationship is likely influenced by other factors in the model. For example, when the magnitude of the sentence is greater, judges may be perceived as less influential, particularly in cases in which the evidence is strong toward guilt.¹⁷³

A. Cumulative Relationships in the Model

Multiple regression analyses are used next to explore combinations of variables in the model.¹⁷⁴ The regression analyses employ several of the variables as predictors of either trial outcome or

¹⁶⁷ $r = .60$, $p = .12$, $n = 8$ (for hand movements); $r = .84$, $p = .009$, $n = 8$ (for judicial); $r = .84$, $p = .009$, $n = 8$ (for directive); $r = .89$, $p = .003$, $n = 8$ (for confident).

¹⁶⁸ $r = .86$, $p = .007$, $n = 8$ (for smiles), and same trend for eye contact ($r = .57$, $p = .14$, $n = 8$).

¹⁶⁹ The potential impact on trial outcome of the dynamics of appearance of justice in jury as compared to bench trials presents an interesting question for future study. See, e.g., Empirical Research, supra note 1, at 800-01; see infra notes 256-60 and accompanying text (discussing error). Note that the outcome-sentence ("D-F") relationships are not tested in this Article (all guilty verdicts resulted in some sentence). Further analysis is needed to describe how guilty verdicts in these bench trials relate to the particular sentence imposed by the judge. As suggested above, this simple relationship is likely affected by other variables in the model (e.g., defendant's criminal history) in meaningful ways. Interestingly, analysis of this simple relationship also may reflect judges' sentencing patterns or their adherence to legislatively mandated sentencing guidelines.

¹⁷⁰ For judges, $r = -.22$, $p = .12$, $n = 52$ (influence); $r = .16$, $p = .27$, $n = 52$ (competence). For prosecution counsel, $r = .37$, $p = .007$, $n = 52$ (competence). For defense counsel, $r = -.87$, $p = .005$, $n = 8$ (influence).

¹⁷¹ As suggested earlier, it is also possible that when the evidence is stronger toward guilt and the result is a guilty verdict, judges are perceived as less influential in the trial process. The cumulative and partial correlation tests in Section B below shed further light on this issue. For example, note the partial correlation finding that when controlling for strength of evidence, judges are rated as more competent when the verdict is guilty. See infra text accompanying note 198.

¹⁷² $r = .22$, $p = .17$, $n = 40$ (suggesting a weak trend for influence). No other results emerged for this relationship.

¹⁷³ This "C'-D-G predicting F" chain will be explored in future tests of the model. Cf. infra notes 186-93 and accompanying text (discussing regression tests of model).

¹⁷⁴ The regression analyses enable a detailed assessment of the relationship between a set of variables and one other variable. For a review of multiple regression techniques, see COHEN & COHEN, supra note 100, at 7 (explaining that multiple regression analyses describe relationships between a complex set of predictor variables and a single criterion variable).

magnitude of sentence imposed.¹⁷⁵ The analyses identify the independent and additive effects of the strength of the evidence and extralegal forces on trial outcome and sentencing.

1. Preliminary Tests

The regression analyses use several combinations of the background, expectancy, communicative, evidentiary, and competence measures as independent predictor variables. The two dependent measures are trial outcome and magnitude of the sentence.¹⁷⁶

Of course, many combinations of the measures in the model may be employed as variables in the regression equations.¹⁷⁷ A positive relationship between any independent variable and the dependent variable indicates a higher rating on that measure when the verdict was guilty or when the magnitude of the sentence was greater. A negative relationship indicates the converse.

a. Example 1: "A-B-C-C'-G" Predicting "D"

This model uses trial outcome as the dependent measure and the following as independent predictors: (1) defendants' gender, (2) ratings of the judges' expectations for trial outcome, (3) ratings of the judges' seven micro behaviors, (4) ratings of the strength of the evidence, and (5) ratings of judges' and prosecution counsels' competence.¹⁷⁸ The R2 for this regression equation is statistically significant: $R^2 = .575$, $F(12, 39) = 4.40$, $p < .0002$. The individual results of this test of the model are shown in Table 1.¹⁷⁹

TABLE 1

¹⁷⁵ For the test of the model in the Stanford Study of the extent to which trial outcomes are predicted by the set of variables, including: the defendants' criminal histories, judges' expectations for trial outcome, the engaged micro behaviors of the judge at trial, and the magnitude of judge-jury agreement, see supra note 101 and accompanying text.

¹⁷⁶ For an explanation of the analyses and statistical testing, see supra note 74. See also COHEN & COHEN, supra note 100, at 49-50, 104. All tests of statistical significance are "two-tailed." The term "ns" refers to the result being statistically not significant at the $p \leq .10$ level.

¹⁷⁷ At some point, the inclusion of too many independent variables in the regression model inhibits meaningful interpretation of the results. Therefore, selected regression models of theoretical interest are presented.

¹⁷⁸ The sample size for ratings of defense counsel was not large enough to justify inclusion of this measure in the regression analyses. See supra note 112.

¹⁷⁹ The regression coefficients in Tables 1-4 are used to form a linear combination of independent variables to estimate the dependent variable (e.g., verdict or magnitude of sentence). The explained variance for each independent variable represents the contribution of each variable in the model, controlling for the effects of all other variables. See COHEN & COHEN, supra note 100, at 39-40. For the individual variables in the model, a conservative estimate of explained variance is computed by dividing the sum of squares (Type II in SAS computer program) by the corrected total sum of squares. Type II sum of squares reflects the variance accounted for assuming that variable is entered last in the regression equation. The combined individual explained variance will be less than the total explained variance.

Logistic regression is also an appropriate means of assessing relationships between combinations of independent measures with a dichotomous dependent measure, such as verdict. However, the sample size of the Iowa Study was insufficient for a logistic regression employing the independent variables found in Tables 1 and 2. Hence, only results from least-squares regression are presented. However, a stepwise logistic regression was used to select the most substantial variables. When selecting from the variables in Table 1, the following were considered: strength of evidence ($b = 5.25$, Chi Square = 3.39, $p = .07$), prosecution competence ($b = 4.56$, Chi Square = 3.71, $p = .06$), and self-touching ($b = 3.93$, Chi Square = 3.16, $p = .08$). A Chi Square test based on the log-likelihood function was significant for the model (Chi Square = 43.68, $df = 3$, $p = .0001$). When selecting from the variables in Table 2, the following were considered: strength of evidence ($b = 5.14$, Chi Square = 2.44, $p = .12$), prosecution competence ($b = 7.61$, Chi Square = 1.91, $p = .17$), and judicial global style ($b = 5.70$, Chi Square = 1.83, $p = .18$). A Chi Square test based on the log-likelihood function was significant for this model (Chi Square = 45.46, $df = 3$, $p = .0001$). These exploratory analyses indicate that future studies should include the above-mentioned and other variables as increased sample size allows.

TEST OF THE MODEL: EXAMPLE 1

<u>Variable</u>	<u>Regression Coefficient</u>	<u>t-value</u>	<u>p-value</u>	<u>Explained Variance</u>
“A” Defendant Gender [†]	.076	.639	.53	0.4%
“B” Expectancy [‡]	-.000	-.007	.99	0.0%
“C” Micro Behaviors: [*]				
Eye Contact	-.056	-.955	.35	1.0%
Smiles	-1.340	-1.660	.11	3.0%
Head Nods	-.004	-.031	.98	0.0%
Forward Leans	.073	.242	.81	0.1%
Hand Movements	.075	1.131	.27	1.4%
Postural Changes	.083	.293	.78	0.1%
Self-Touching	-.044	-.487	.63	0.3%
“C” Strength of Evidence [?]	.127	3.841	.0004 ^{††}	16.0%
“G” Competence: ^{‡‡}				
Judge	-.118	-1.679	.11	3.1%
Prosecution	.127	2.237	.04 ^{††}	5.4%

Total Explained Variance = 58%

Table 1 shows that several variables in the model together and independently predict trial outcomes. Guilty verdicts are predicted when: (1) judges display fewer smiles throughout the trial, (2) the strength of the evidence is rated as strong toward guilt, (3) the judge is rated as less competent, and (4) the prosecution is rated as more competent.¹⁸⁰

Neither defendant gender nor judges' expectations for trial outcomes independently add to the prediction of trial outcomes. In this test of the model, the strongest independent predictor of trial outcomes is strength of the evidence (explaining 16% of the variance).¹⁸¹ The combined explained variance of the micro behaviors is approximately 6%, suggesting that these variables add independently (albeit modestly) to the prediction of trial outcomes.¹⁸²

[†] Coded as 1 = men, 2 = women.

[‡] Coded from 1 to 9; 1 = not guilty, 9 = guilty

^{*} Coded by counting, from less to more.

[?] Coded from 1 to 9; 1 = not guilty, 9 = guilty.

^{††} Statistical significance assessed at $p \leq .10$.

^{‡‡} Coded from 1 to 9; 1 = not competent, 9 = very competent.

¹⁸⁰ Analogous ratings are not available for defense counsel as in most cases defendants represented themselves. See supra note 112 and accompanying text.

¹⁸¹ The R² for strength of the evidence variable only is .462, showing that addition of defendant gender, expectancy, micro behaviors, and competence result in an increase of .114 (differences not significant, $F = .95$, $p = .51$, $df = 11, 39$).

¹⁸² When the regression is conducted, substituting defendant age for gender, the $R^2 = .585$, $F(12, 39) = 4.59$, $p < .0001$. The significant independent predictor variables are: (1) strength of the evidence, $t = 3.68$, $p = .0007$; and (2) prosecution competence, $t = 2.03$, $p = .05$.

a. Example 2: "A-B-C-C'-G" Predicting "D"

The next test of the model uses the same independent predictors from Example 1 above but substitutes the global "C" variables for the micro variables.¹⁸³ The dependent measure remains trial outcome. The R2 for the regression equation is statistically significant: R2 = .601, F(9, 42) = 7.03, p < .0001. The individual results of this test are shown in Table 2.

TABLE 2

TEST OF THE MODEL: EXAMPLE 2

<i>Variable</i>	<i>Regression Coefficient</i>	<i>t-value</i>	<i>p-value</i>	<i>Explained Variance</i>
"A" Defendant Gender [†]	-.003	-.031	.98	0.0%
"B" Expectancy [‡]	-.018	-.358	.73	0.1%
"C" Global Behaviors: [*]				
Judicial	-.092	-.710	.49	.05%
Warm	.084	1.678	.10 [?]	2.7%
Directive	-.60	-1.137	.27	1.2%
Confident	-.020	-.170	.87	0.0%
"C" Strength of Evidence [?]	.145	4.657	.0001 ^{**}	20.6%
"G" Competence: ^{‡‡}				
Judge	-.056	-.577	.57	0.3%
Prosecution	.141	2.579	.02 ^{**}	6.3%

Total Explained Variance = 60%

Table 2 shows that several variables in the model together and independently predict trial outcomes. Guilty verdicts result when: (1) judges are warmer throughout trials (i.e., reflect a supportive appearance), (2) the strength of the evidence is rated as strong toward guilt, and (3) the prosecution is rated as more competent.

Neither defendants' gender nor judges' expectations for trial outcomes add to the prediction of trial outcomes. The strongest independent predictor of trial outcome is again the strength of the

¹⁸³ Again, for these exploratory tests, the goal is to limit the number of variables tested in the model to aid in its preliminary interpretation.

[†] Coded as 1 = men, 2 = women.

[‡] Coded from 1 to 9; 1 = not guilty, 9 = guilty

^{*} Coded by counting, from less to more.

[?] Statistical significance assessed at p ≤ .10.

[?] Coded from 1 to 9; 1 = not guilty, 9 = guilty.

^{**} Statistical significance assessed at p ≤ .10.

^{‡‡} Coded from 1 to 9; 1 = not competent, 9 = very competent.

evidence (explained variance = 20.6%).¹⁸⁴ An independent effect of judges' warmth on trial outcomes also appears (explained variance = 2.7%).¹⁸⁵ The global styles contribute modestly to the prediction of outcome (separately, explained variance=4.4%; together, explained variance = 8%).

a. Example 3: "A-B-C-C'-G" Predicting "F"

This test employs the magnitude of the sentence as the dependent measure.¹⁸⁶ The independent predictors are: (1) defendants' gender, (2) ratings of the judges' expectations for trial outcome, (3) ratings of the judges' seven micro communicative behaviors, (4) ratings of the strength of the evidence, and (5) ratings of judges' and prosecution counsels' competence. Consistent with the pattern of simple relationships, the R2 for the regression equation is not statistically significant: R2 = .304, F(12, 27) = .981, p = .49. This nonfinding suggests that, when considering all the variables in this model together, these judges are not influenced substantially by these factors in their sentencing decisions.¹⁸⁷

The individual results of this test of the model are shown in Table 3.

TABLE 3
TEST OF THE MODEL: EXAMPLE 3

<i>Variable</i>	<i>Regression Coefficient</i>	<i>t-value</i>	<i>p-value</i>	<i>Explained Variance</i>
"A" Defendant Gender [†]	-.244	-1.671	.11	7.2%
"B" Expectancy [‡]	.051	.673	.51	1.2%
"C" Micro Behaviors: [*]				
Eye Contact	.063	.712	.49	1.3%
Smiles	.235	.234	.82	0.1%
Head Nods	.255	1.545	.14	6.2%
Forward Leans	.038	.099	.93	0.0%
Hand Movements	-.089	-1.170	.26	3.5%
Postural Changes	-.562	-1.593	.13	6.5%
Self-Touching	-.093	-.689	.50	1.2%

¹⁸⁴ The R2 for strength of the evidence variable only is .462, showing that the addition of defendant gender, expectancy, global behaviors, and competence result in an increase of .139 (difference is significant, F = 1.83, p = .10, df (8, 42)).

¹⁸⁵ When the regression is conducted substituting defendant age for gender, the R2 = .607, F = 7.20, p < .0001, df (9, 42). The significant independent predictor variables are: strength of the evidence, t = 4.45, p = .0001; and prosecution competence, t = 2.48, p = .02.

¹⁸⁶ The sample size for this test of the model is 39 cases (i.e., only those cases in which a guilty verdict was found).

¹⁸⁷ This is not to suggest that the same 30% of the variance explained by the model is, in practical terms, insignificant. See Robert Rosenthal & Peter D. **Blanck**, Science and Ethics in Conducting, Analyzing, and Reporting Social Science Research: Implications for Social Scientists, Judges, and Lawyers, 68 IND. L.J. 1209 (1993).

[†] Coded as 1 = men, 2 = women.

[‡] Coded from 1 to 9; 1 = not guilty, 9 = guilty

^{*} Coded by counting, from less to more.

“C” Strength of Evidence [?]	-0.31	-0.648	.53	1.1%
“G” Competence: ^{‡‡}				
Judge	.159	1.677	.11	7.3%
Prosecution	-.125	-1.594	.13	6.6%

Total Explained Variance = 30%

Table 3 shows that the variables in this regression, when controlling for the effects of the other variables, do not predict judges' sentencing patterns. The following measures alone tend to relate to the imposition of relatively more severe sentences: (1) cases involving defendants who are men, and (2) cases in which the judge is rated as more competent.¹⁸⁸

These findings echo those for the simple relationships, suggesting that some individual variables may alone, but not in combination, predict sentencing patterns of these judges.¹⁸⁹ Strikingly, strength of the evidence explains only 1.1% of the variance in predicting sentences.¹⁹⁰ This finding contrasts dramatically with the substantial effect of strength of the evidence in predicting trial outcomes. Likewise, the individual combination of the micro behaviors explains roughly 19% of the variance in predicting sentences. This finding is also dramatic when compared to the combined impact of the micro behaviors when predicting trial outcomes (e.g., explaining 6%).

d. Example 4: "A-B-C-C'-G" Predicting "F" The final test substitutes the global variables for the micro variables but retains as the dependent measure the magnitude of the sentence.¹⁹¹ The R2 for this regression equation is: R2 = .327, F(9, 30) = 1.62, p = .16. The individual results of this test are shown in Table 4.

TABLE 4

TEST OF THE MODEL: EXAMPLE 4

<i>Variable</i>	<i>Regression Coefficient</i>	<i>t-value</i>	<i>p-value</i>	<i>Explained Variance</i>
“A” Defendant Gender [†]	-.41	-.914	.37	1.8%
“B” Expectancy [‡]	1.06	1.547	.14	5.4%
“C” Global Behaviors: [*]				
Judicial	-.224	-1.433	.17	4.6%

[?] Coded from 1 to 9; 1 = not guilty, 9 = guilty.

^{‡‡} Coded from 1 to 9; 1 = not competent, 9 = very competent.

¹⁸⁸ Analogous ratings are not available for defense counsel. See supra note 112 and accompanying text.

¹⁸⁹ When the regression is conducted, substituting defendant age for gender, R2 = .232, F(12, 27) = 0.681, p = .76. The only independent predictor variable is nods: t = 1.65, p = .12 (suggesting a trend only).

¹⁹⁰ The R2 for strength of evidence variable only is .0001, showing that the addition of defendant gender, expectancy, micro behaviors, and competence result in an increase of .304 (difference is not significant, F = 1.07, p = .42, df (11, 27)).

¹⁹¹ The sample size for this test of the model is 39 cases (e., only those cases in which a guilty verdict was found).

[†] Coded as 1 = men, 2 = women.

[‡] Coded from 1 to 9; 1 = not guilty, 9 = guilty

^{*} Coded by counting, from less to more.

Warm	.093	1.437	.17	4.6%
Directive	.058	.783	1.4%	
Confident	.184	1.148	.26	3.0%
“C” Strength of Evidence [?]	-.027	-.590	.56	0.8%
“G” Competence: ^{‡‡}				
Judge	.103	.837	.41	1.6%
Prosecution	-.136	-1.646	.11	6.1%

Total Explained Variance = 33%

Table 4 indicates that the variables in this regression, when controlling for the effects of the other variables, do not well predict judges' sentencing.¹⁹² In this test of the model, the strength of the evidence alone explains less than 1% of the variance in predicting sentences (as compared to between 16% and 21% for predicting trial outcomes). In contrast, the global styles explain 14% of the variance in predicting sentences (as compared to 4% for predicting trial outcomes).¹⁹³

The findings of the four sample regression equations illustrate the extent to which several combinations of measures are useful for predicting trial outcomes and sentence magnitudes.¹⁹⁴ Despite the strong independent effect of the strength of the evidence, extralegal factors, such as judges' communicative variables, emerge as modest independent predictors of trial outcomes. The same combinations of variables do not strongly predict sentencing patterns.

Taken together, the findings of the regression analyses suggest three preliminary conclusions: (1) strength of the evidence is a strong predictor of trial outcomes but not of sentencing patterns; (2) judges' communicative behavior during trial is a relatively better predictor of sentencing patterns than of trial outcomes; and (3) in bench trials, judges appear able to separate factors affecting their fact finding role (i.e., the determination of guilt) from those affecting their sentencing function.¹⁹⁵

Further research with actual trials is warranted to understand the relative impact of legal and extralegal factors on judge decision making.¹⁹⁶ The next section uses partial correlational

[?] Coded from 1 to 9; 1 = not guilty, 9 = guilty.

^{‡‡} Coded from 1 to 9; 1 = not competent, 9 = very competent.

¹⁹² When the regression is conducted, substituting defendant age for gender, $R^2 = .325$, $F(9, 30) = 1.61$, $p = .16$. Interestingly, several independent predictor variables emerge: (1) judges' expectancy for guilty, $t = 1.86$, $p = .08$; (2) warm style, $t = 1.72$, $p = .10$; and (3) prosecution competence, $t = -1.73$, $p = .10$. Further examination of this combination of variables may be warranted.

¹⁹³ The R^2 for strength of evidence variable only is .0001, showing that the addition of defendant gender, expectancy, global behaviors, and competence result in an increase of .327 (indicating that difference is marginally substantial, $F = 1.82$, $p = .12$, $df(8, 30)$).

¹⁹⁴ That sentencing decisions were not substantially predicted by the variables in the model is an important result. As discussed in the final Part of this Article, the study of the factors that influence the sentencing process is a complicated task that calls for further research in its own right. See *infra* notes 274-88 and accompanying text.

¹⁹⁵ Cf. Kevin R. Reitz, *Sentencing Facts: Travesties of Real- Offense Sentencing*, 45 STAN. L. REV. 523, 573 (1993) (regarding need for theory to segregate conviction issues from sentencing issues).

¹⁹⁶ In subsequent tests of the model, the independent effect of the "A" variable ("seriousness of the offense") will be explored. Seriousness of the offense is coded in the Iowa Study by the maximum possible sentence for a particular charge. Preliminary bivariate analyses show that seriousness of the offense relates to: (1) more leans ($r = .30$, $p = .06$, $n = 40$) and postural changes ($r = .36$, $p = .02$, $n = 40$) by the judge during the trial, but (2) less warm behavior by the judge during the trial ($r = -.42$, $p = .006$, n

analyses to explore in greater detail the independent effect of the strength of the evidence on trial outcomes and sentencing.

2. Partial Correlations Controlling for Strength of Evidence

Partial correlations describe the relationship between any independent variable and a dependent variable, statistically controlling for the effects of other variables in the model.¹⁹⁷ Partial correlation analyses here explore the extent to which predictions of trial outcome and sentence magnitude are based primarily on an independent relationship with the strength of the evidence variable.

Table 5 presents the partial correlations between trial outcomes and defendants' backgrounds, judges' expectations, judges' communicative variables, and competence measures, controlling for the strength of the evidence.¹⁹⁸

Examination of the partial correlations between verdict and other variables in the model reveals several findings. Foremost, controlling for the strength of the evidence, two of the judges' communicative variables alone tend to predict guilty verdicts: (1) fewer smiles and (2) less judicial behavior. Second, taking into account the strength of the evidence, there is a trend for older defendants to receive more guilty verdicts than younger defendants. Third, when controlling for the strength of the evidence, prosecution counsel are rated as more competent when the verdict is guilty. The results of the partial correlations highlight the suggestion that, although the strength of the evidence may be a central factor in judge or jury decision making, other extralegal factors independently influence this process in significant ways. This extralegal influence may be particularly apparent in cases in which the evidence is close.¹⁹⁹ This conclusion is consistent with the Stanford results.²⁰⁰

= 40). Future analyses will explore the seriousness of the offense as an independent predictor of trial outcome and magnitude of the sentence.

¹⁹⁷ See COHEN & COHEN, *supra* note 100, at 83, 181-82 (explaining that partial correlation is the relationship between two variables with other independent variables held constant).

¹⁹⁸ The *df* for this statistical test is 49. Analogous tests are performed controlling for the complexity of the evidence but are not reported here.

¹⁹⁹ Close cases are artificially defined here by using partial correlation analyses to control for the strength of the evidence.

²⁰⁰ See *supra* note 98 and accompanying text.

TABLE 5
PARTIAL CORRELATIONS WITH VERDICT ('D'):
CONTROLLING FOR STRENGTH OF THE EVIDENCE ('C')

<i>Variable</i>	<i>Partial Correlation</i>	<i>p-value</i>
"A" Background: [†]		
Defendant Age	.181	.11
Defendant Gender	.005	.49
"B" Expectancy		
	.128	.19
"C" Communicative:		
Eye Contact	-.074	.30
Smiles	-.231	.06 [‡]
Head Nods	.004	.49
Forward Leans	-.054	.36
Hand Movements	.077	.30
Postural Changes	-.135	.17
Self-Touching	-.066	.33
Judicial		
Warm	-.199	.09 [‡]
Directive	.126	.19
Confident	-.124	.20
	-.169	.12
"G" Competence:		
Judge	-.063	.34
Prosecution	.230	.06 [‡]
Defense*	-.263	.29

Table 6 presents the partial correlations between magnitude of sentence, defendants' backgrounds, judges' expectations, judges' communicative variables, and competence measures, controlling for the strength of the evidence.²⁰¹

Examination of the partial correlations between the magnitude of the sentence with the other independent measures in the model, controlling for the strength of the evidence, reveals that two communicative variables independently tend to predict the magnitude of the sentence: (1) fewer postural changes, yet (2) warmer behavior (e.g., reflecting a post-verdict attempt to appear just toward the defendant).

The preliminary findings of the Iowa Study suggest several directions for future inquiry. The results are encouraging, given that observers were able to assess reliably many trial factors that

[†] Partial correlation for defendant socioeconomic status is -.014, $p = .46$.

[‡] Statistical significance assessed at $p \leq .10$.

* $n=8$.

²⁰¹ The df for this statistical test is 37.

were then shown to relate to judges' behavior, decision making, and sentencing. The final Part of this Article explores the implications of the findings, with emphasis on the need for future empirical study of the appearance of justice, trial error, and sentencing patterns.

TABLE 6

PARTIAL CORRELATIONS WITH MAGNITUDE OF SENTENCE ('F'):
CONTROLLING FOR STRENGTH OF THE EVIDENCE ('C')

<i>Variable</i>	<i>Partial Correlation</i>	<i>p-value</i>
"A" Defendant Gender:		
Defendant Age	-.061	.36
Defendant Gender	-.195	.12
"B" Expectancy	.172	.15
"C" Communicative		
Eye Contact	-.022	.45
Smiles	-.137	.21
Head Nods	.197	.12
Forward Leans	-.096	.28
Hand Movements	-.074	.33
Postural Changes	-.231	.08 [†]
Self-Touching	.059	.36
Judicial	.029	.44
Warm	.332	.02 [†]
Directive	.146	.19
.113	.25	
"G" Competence:		
Judge	.124	.23
Prosecution	-.405	.41
Defense [‡]	-.826	.19

V. EMERGING ISSUES IN THE STUDY OF TRIAL PROCESSES

This Article has set forth a model for exploring empirically the impact of legal and extralegal factors on trial outcomes and sentencing. There is much to be learned about how these factors together and independently influence trial decision making.²⁰² The developing empirical research

[†] Statistical significance assessed at $p \leq .10$.

[‡] $n = 4$.

²⁰² Communicating with Juries, Panel Two: Innovations for Improving Courtroom Communication and Views from Appellate Courts, The Annenberg Washington Program Conference, April 10, 1992, 68 IND. L.J. 1061, 1067 (1993) [hereinafter Panel Two] (statement of Judge Patricia Wald) (commenting that there is little known about the impact of trial processes on trial outcomes).

cannot yet inform researchers, courts, and practitioners about many of the issues concerning the complexity of judges' behavior and decision making.²⁰³ This is true because of the constraints of sampling, the inability of researchers to assign defendants randomly to bench or jury trials, and ethical considerations involving research with actual trials.²⁰⁴ [FN204]

Many judges at the trial and appellate levels remain unaware of the impact of extralegal forces on trial outcomes or of the developing research on trial processes.²⁰⁵ [FN205] This final Part raises several emerging issues in the development of research on judges' behavior.

A. Toward a More Cumulative View of Social Science Research

Researchers, courts, and legal practitioners often raise questions about the generalizability of actual courtroom behavior studies and the extent to which the results of any particular study would hold true across the population of judges or over different types of trials.²⁰⁶ [FN206] The primary means for addressing these questions lie in the replication and refinement of the results of any study for legal scholars, practitioners, and courts to develop a more cumulative view of social science research.

This Article has presented information derived from an ongoing effort to describe the impact of various factors on trial processes and outcomes in bench trials. The research uses a theoretical model for exploring these issues. By building on the Stanford Study, the extended test of the model in Iowa bench trials identified patterns of findings that not only have implications for this set of trials conducted by these judges, but for courtroom processes conducted by other judges in other types of cases as well.

In the Stanford Study, the magnitude of the findings for the potential impact of trial judges' nonverbal behavior alone was about 9% of the variance, or in correlational terms, on average a correlation of approximately 0.3.²⁰⁷ [FN207] This is a substantial effect by many research standards, but it leaves a good amount of variation in the results unexplained.²⁰⁸ [FN208] Still, this effect may be especially noteworthy to counsel of the 9% of defendants found guilty or not guilty as a result of the judges' communicative behavior.²⁰⁹ [FN209]

²⁰³ See generally STEPHEN A. SALTZBURG, AM. BAR ASS'N & BROOKINGS INST., IMPROVING THE QUALITY OF JURY DECISION MAKING: EXECUTIVE SUMMARY, SYMPOSIUM ON THE FUTURE OF THE CIVIL JURY SYSTEM IN THE UNITED STATES (1992).

²⁰⁴ For a review, see INTERPERSONAL EXPECTATIONS, *supra* note 41, at 339-43 (exploring the "process" of research in the courtroom). See also Rosenthal & **Blanck**, *supra* note 187; Hagan & Bumiller, *supra* note 6, at 10-18 (noting other methodological issues).

²⁰⁵ Panel Two, *supra* note 202, at 1067 (statement of Judge Patricia Wald). Other appellate judges are skeptical generally of the value of juries as fact finders. *Id.* at 1067 (statement of Justice James D. Heiple) ("I am very skeptical of the value of juries and probably, if it were up to me, would abolish juries in civil cases.").

²⁰⁶ Additional study is needed of pro se trials versus cases in which the defendant is represented by counsel. This study is especially warranted in light of the case law that pro se defendants be accorded their constitutional right to a fair trial. See *Oses v. Massachusetts*, 775 F. Supp. 443, 456, 465, (D. Mass. 1991) (citing U.S. Supreme Court cases on the rights of the pro se defendant, and noting that the trial judge "polluted the atmosphere [of the trial] with his own actions and his inability properly to control the actions of the other trial participants").

²⁰⁷ Panel One, *supra* note 3, at 1056 (statement of Robert Rosenthal).

²⁰⁸ See Rosenthal, *supra* note 5, at 775-76 (providing other examples of studies showing small effect sizes, yet having dramatic practical importance).

²⁰⁹ See Panel One, *supra* note 3, at 1056 (statement of Robert Rosenthal) (regarding effect of Stanford Study).

In the Iowa Study, preliminary regression analyses show the predictive power of the model. When taking into account various combinations of variables to predict trial outcomes, the model predicts 58% to 60% of the explained variance, or in correlational terms, an average correlation of approximately 0.76.²¹⁰ [FN210] This substantial finding illustrates the practical importance of the model, while at the same time acknowledging that bench trials are complex interactions that involve many other factors.²¹¹ [FN211]

The Iowa Study also shows that extralegal forces independently influence trial outcomes in significant ways.²¹² [FN212] For instance, the micro behaviors alone account for approximately 6% and the global styles alone for approximately 4% of the explained variance. These results support the general magnitude of the effect found in the Stanford Study.

Additionally, in the Iowa Study, the strength of the evidence independently accounts for some 16% to 21% of the explained variance in predicting trial outcomes. Likewise, the competence ratings of the judge and prosecution counsel add to the predictive power of the model, together contributing approximately 7% to 8% of the explained variance in predicting trial outcomes.

Examination of partial correlations further supports the substantial relation between trial outcomes and extralegal variables. These findings illustrate that, when controlling for the strength of the evidence, extralegal forces independently predict trial outcomes and sentencing. Thus, this influence may be particularly apparent in close cases. Had the Iowa Study not been performed, little information would exist about the relative impact in bench trials of evidentiary and extralegal factors on trial outcomes and sentencing.²¹³ [FN213]

The Iowa Study also shows that combinations of variables in the model do not significantly predict judges' sentencing patterns. Though not unsubstantial in magnitude, in contrast to the results for trial outcomes, the total explained variance for sentencing ranges from 30% to 33%. Most striking is the finding that the strength of the evidence variable independently accounts for only 1% of the variance in the prediction of sentence magnitude. This finding is dramatic when compared to the substantial effect of this variable on trial outcomes. In addition, the explained variance of the independent effect of the micro behaviors and global styles ranges from 14% to 19% in predicting the magnitude of sentences. Taking into account the evidence of the case, however, sentencing decisions still are not related substantially to extralegal variables in the model.²¹⁴ [FN214]

²¹⁰ See supra note 179 and accompanying text (reporting results of regression equation for Example 1, and predicting trial outcome).

²¹¹ Lawrence P. Tiffany et al., A Statistical Analysis of Sentencing in Federal Courts: Defendants Convicted After Trial, 4 J. LEGAL STUD. 369 (1975) (noting that in federal courts, conviction by jury generally leads to a more severe sentence than conviction in bench trials).

²¹² Galen V. Bodenhausen, Second Guessing the Jury: Stereotypes and Hindsight Biases in Perceptions of Court Cases, 20 J. APPLIED SOC. PSYCHOL. 1112, 1118 (1990) (reporting results of simulated study showing that expectations held by mock jurors had powerful effect on subsequent decision making).

²¹³ For a detailed empirical study of bench trials, see Stephen J. Schulhofer, Is Plea Bargaining Inevitable?, 97 HARV. L. REV. 1037, 1067, 1107 (1984) (concluding that formal bench trial permits expeditious, fair, and accurate resolution of criminal cases, and noting differences in style and demeanor of public defenders in bench trials).

²¹⁴ Compare id. at 1070 (noting judges' active "judicial behavior" in bench trials) with supra notes 60-73 and accompanying text (describing judges' global directive styles).

Legal scholars and practitioners typically have exaggerated the impact of extralegal variables on trial outcomes and underestimated the extent to which judge or juror decision making is affected by the strength and quality of the evidence.²¹⁵ [FN215] Quite to the contrary, the Iowa findings show that the strength of the evidence is generally a better predictor of trial outcomes than are extralegal factors. On the other hand, the "appearance of justice" is a relatively better predictor of sentencing.

Without further replication, the Stanford Study and the Iowa Study must continue to be interpreted with caution, despite the predicted pattern of findings.²¹⁶ [FN216] The findings do not imply necessarily that similar effects will always be found in other trials and courts and with other judges.²¹⁷ [FN217] The present program of research illustrates, however, a view of social science research as a cumulative endeavor.²¹⁸ [FN218] Courts, legal scholars, and practitioners need to become increasingly familiar with the emerging statistical techniques for assessing this body of social science research.²¹⁹ [FN219] Future research will refine the preliminary findings and further evaluate the magnitude of the effects found here.²²⁰ [FN220] Research will also need to be conducted with actual and mock trial participants to reveal the effects of legal and extralegal forces on judges' decision making.²²¹ [FN221]

A cumulative view of social science study also enables courts and practitioners to better evaluate the practical importance of the present body of research.²²² [FN222] Less attention is focused on particular results and more on the overall magnitude of the phenomenon in question.²²³ [FN223] Likewise, less emphasis is placed on the results of a single study as determinative of a particular

²¹⁵ Tanford & Tanford, *supra* note 4, at 755 (concluding that simulated trial studies that manipulate evidentiary strength show that extralegal factors are most powerful when trial evidence is weak).

It will be interesting to examine also the extent to which different trial participants—judges, jurors, lawyers—estimate the effect that legal and extralegal factors have on trial outcomes. Cf. Lee E. Teitelbaum et al., *Evaluating the Prejudicial Effect of Evidence: Can Judges Identify the Impact of Improper Evidence on Juries?*, 1983 WIS. L. REV. 1147, 1172 (reporting results that prosecutors and defense attorneys view evidence differently). As suggested above, lawyers may tend to exaggerate the probable effect that extralegal factors have on outcome and to ignore the fact that juror decision making is affected primarily by the strength of the evidence. Tanford & Tanford, *supra* note 4, at 755 (summarizing studies of mock trials showing that extralegal factors have their greatest impact when the evidence is strong and that the relative impact of extralegal factors is small compared to the impact of evidentiary factors).

²¹⁶ For a review, see Tanford & Tanford, *supra* note 4, at 771-79 (arguing the benefits of psychologist-lawyer collaboration).

²¹⁷ *Id.* at 755, 771-79.

²¹⁸ See Rosenthal, *supra* note 5 (providing overview of this trend); Ralph L. Rosnow & Robert Rosenthal, *Statistical Procedures and the Justification of Knowledge in Psychological Science*, 44 AM. PSYCHOLOGIST 1276 (1989) (same).

²¹⁹ See Michael J. Saks & Peter D. Blanck, *Justice Improved: The Unrecognized Benefits of Aggregation and Sampling in the Trial of Mass Torts*, 44 STAN. L. REV. 815, 845-47 (1992) (arguing that the courts need to become familiar with aggregation techniques such as meta-analysis).

²²⁰ This view is furthered by the development of theoretical models, such as the one presented herein, that aid in the precision of replication of research studies. See Robert Rosenthal, *Cumulating Psychology: An Appreciation of Donald T. Campbell*, 2 PSYCHOL. SCI. 213, 217, 219 (1991) (noting the importance of replication and a more cumulative view of social science research through meta-analytic statistical techniques); see also Peter D. Blanck, *Empirical Study of the Employment Provisions of the Americans with Disabilities Act: Methods, Preliminary Findings, and Implications*, 22 N.M. L. REV. 121 (1992) (noting the importance of not overgeneralizing from the results of a single social science study); Peter D. Blanck, *The Emerging Work Force: Empirical Study of the Americans With Disabilities Act*, 16 J. CORP. L. 693 (1991) (same).

²²¹ Tanford & Tanford, *supra* note 4, at 755 (noting the importance of understanding the relative effect of legal and extralegal factors on trial outcomes).

²²² Cf. Panel One, *supra* note 3, at 1057 (statement of Judge LaDoris Cordell) (noting that California judges have not incorporated learning from study into training programs nor have they tried to replicate findings).

²²³ Rosnow & Rosenthal, *supra* note 218, at 1278-79 (noting the importance of defining the results of social science research in effect-size terms).

legal fact.²²⁴ [FN224] The importance of cumulation to the interpretation of social science research requires further discussion than allotted here.²²⁵ [FN225]

B. Using the Model to Assess Trial Error

Appellate courts evaluate the propriety of legal and extralegal factors in judge and jury decision making.²²⁶ [FN226] The model enables appellate courts to review more systematically alleged error in criminal trials. It may also further the understanding of the extent to which the alleged error is harmless or prejudicial to the rights of the criminal defendant.²²⁷ [FN227]

Under Federal Rule of Criminal Procedure 52(a), harmless error is "[a]ny error, defect, irregularity or variance which does not affect substantial rights"²²⁸ [FN228] In *Chapman v. California*,²²⁹ [FN229] the Supreme Court held that a lower court must find "beyond a reasonable doubt that the error complained of did not contribute to the verdict obtained"²³⁰ [FN230] before federal constitutional error can be harmless. In essence, the Court found that there may be some constitutional errors that do not require automatic reversal.²³¹ [FN231]

The Court also recognized in *Chapman* that some "constitutional rights [are] so basic to a fair trial that their infraction can never be treated as harmless."²³² [FN232] One such right enunciated by the Court is the guarantee to a trial before an impartial judge.²³³ [FN233] The Court affirmed this position in *Gray v. Mississippi*,²³⁴ [FN234] concluding that the violation of some

²²⁴ *Id.* at 1280.

²²⁵ See generally Rosenthal & **Blanck**, *supra* note 187. For example, another area that may be studied from a more cumulative view is the effect of the increasing use of video evidence in trials, an effect dramatically demonstrated in the Rodney King trials. Technological advances will make videotape evidence crucial to legal processes, whether in will documentation, consent issues, or testimonial evidence. In criminal and civil cases, video evidence not only is useful for examining witness demeanor but also for establishing facts. See, e.g., Peter Huber, *Juries and Justice*, *FORBES*, June 8, 1992, at 136. Moreover, appellate review of videotape evidence will likely raise new issues about the nature and evaluation of evidence.

The study of self-awareness and evidentiary presentation at trial will also be an important next step of study. For a comprehensive review, see generally DePaulo, *supra* note 143. Questions such as the following remain: What awareness do judges and trial participants have of their expression of nonverbal behavior? How may lawyers "match" their sending communication skills with the receiving skills of judges and juries? Many argue, though, that the judicial system is reluctant to open itself to scrutiny of self-analysis, particularly as dockets become increasingly overcrowded. Panel One, *supra* note 3, at 1057 (statement of Judge LaDoris Cordell) (noting that the California judiciary has not responded, at least through its judicial college, to study issues related to communication bias in the courtroom).

²²⁶ For a review, see Teitelbaum et al., *supra* note 215, at 1176-80 (discussing problem raised in analysis of harmless error).

²²⁷ The related issue may be stated as follows: "Should reversal of a trial decision in which error has occurred depend upon a determination by the appellate court of the probable effect in fact of the error, or should it turn on the probable accuracy of the verdict reached below?" *Id.* at 1177.

²²⁸ Fed. R. Crim. P. 52(a). That rule also defines "plain error" as one "affecting substantial rights [that] may be noticed although they were not brought to the attention of the court." *Id.*

²²⁹ 386 U.S. 18 (1967).

²³⁰ *Id.* at 24; cf. *United States v. Rhone*, 864 F.2d 832, 835 (D.C. Cir. 1989).

If, in light of all the circumstances—the language of the instructions, the arguments of counsel, and the evidence itself—it is "highly improbable" that the jury convicted on an improper theory, then "technical errors in the instructions are deemed harmless, and we will affirm." If, on the other hand, these factors indicate a "substantial possibility" that the jury convicted on an improper legal theory, then reversal is required.

Id. (quoting *U.S. v. Lemire*, 720 F.2d 1327, 1339, 1343 (D.C. Cir. 1983), cert. denied, 467 U.S. 1226 (1984)).

²³¹ *Chapman*, 386 U.S. at 22.

²³² *Id.* at 23.

²³³ *Id.* at 23 n.8; see *id.* at 52 n.7 (Harlan, J., dissenting) (concluding that a biased judge could never be deemed a harmless error); see also *supra* notes 9-33 and accompanying text (discussing the appearance of justice).

²³⁴ 481 U.S. 648 (1987).

constitutional rights can never be considered harmless error.²³⁵ [FN235] In *Gray*, the Court declared that "the right to an impartial adjudicator, be it judge or jury, is such a right."²³⁶ [FN236] More recently, in *Arizona v. Fulminante*,²³⁷ [FN237] the Court concluded that a biased judge is one example of "structural" error.²³⁸ [FN238] In *Fulminante*, Chief Justice Rehnquist distinguished "trial" versus "structural" errors.²³⁹ [FN239] Trial errors occur "during the presentation of the case to the jury, and ... may therefore be quantitatively assessed in the context of other evidence presented in order to determine whether its admission was harmless beyond a reasonable doubt."²⁴⁰ [FN240] Structural error, in contrast, affects the "entire conduct of the trial from beginning to end."²⁴¹ [FN241]

The model here allows for the calibration of the independent and combined effects of trial and structural error on trial outcomes and sentencing. Such analyses may illustrate "quantitatively" any real world weakness or strength in the trial/structural error distinction drawn by the Court in *Fulminante*.

Additionally, the model is a method to calibrate the essential purpose of a criminal trial; that is, "to decide the factual question of the defendant's guilt or innocence, and [to] promote[] public respect for the criminal process by focusing on the underlying fairness of the trial rather than on the virtually inevitable presence of immaterial error."²⁴² [FN242] Thus, the model helps in the calibration of the magnitude of error,²⁴³ [FN243] and, in subsequent tests, may prove useful for exploring the contention that without such a framework for analysis, a structural error is one "we can never know with any certainty."²⁴⁴ [FN244]

Along these lines, Professor Ogletree writes:

If the essential distinction between "trial" and "structural" errors lies in whether they are susceptible to being quantified and compared with other evidence, then the *Fulminante* majority was wrong to conclude that all errors that occur at trial are sometimes susceptible of measurement and that all those that pervade the trial "from beginning to end" rarely are.²⁴⁵ [FN245]

Ogletree concludes: "Ultimately, the classification of error must rest on some empirical assumptions about whether, in the real world, errors of this kind are susceptible to harmless error analysis and how often they are clearly harmless."²⁴⁶ [FN246] Thus, the model may be used to test empirical assumptions in the Court's trial/structural error distinction, as well as to further the

²³⁵ Id. at 668.

²³⁶ Id.

²³⁷ 111 S. Ct. 1246 (1991).

²³⁸ Id. at 1265. (Rehnquist, C.J., delivering the opinion in part and dissenting in part).

²³⁹ Id.

²⁴⁰ Id. at 1264 (emphasis added).

²⁴¹ Id. at 1265.

²⁴² Id. at 1264 (citations omitted).

²⁴³ But see id. (White, J., dissenting) (accusing the majority of "drawing a meaningless dichotomy between 'trial errors' and 'structural defects' in the trial process"); see also Ogletree, *supra* note 21, at 161-72 (critiquing trial/structural error analysis).

²⁴⁴ Ogletree, *supra* note 21, at 162.

²⁴⁵ Id. at 163 (showing that even with the existence of a biased judge (structural error), there can still be overwhelming evidence to support a conviction beyond a reasonable doubt).

²⁴⁶ Id. at 164.

discussion of how the harmless error rule may be applied to the case of the biased trial judge.²⁴⁷ [FN247]

Another issue is the model's potential for calibrating the overall effect of error on a decision maker (e.g., on a judge or jury). Recently, in *Yates v. Evatt*,²⁴⁸ [FN248] the Court held that to find an error "did not 'contribute' to the ensuing verdict" is not to suggest that a jury may have been "totally unaware of that feature of the trial later held to have been erroneous."²⁴⁹ [FN249] In *Yates*, Justice Souter wrote: "To say that an error did not contribute to the verdict is ... to find that error unimportant in relation to everything else the jury considered on the issue in question, as revealed in the record."²⁵⁰ [FN250] Whether or not some 9% to 15% of the variance explained by judges' behavior in predicting trial outcomes and sentences is "unimportant in relation to everything else the jury consider s " remains an unanswered (but testable) empirical question.²⁵¹ [FN251]

Additionally, the model allows for review of evidentiary strength and information that may not be conveyed by the written trial record alone.²⁵² [FN252] In the Stanford Study, this was accomplished by videotape analyses, and in the Iowa Study, by on-line courtroom ratings. With either mode, the model provides a means for reviewing courts to assess the impact of various legal and extralegal forces on trial decision making "as a whole."²⁵³ [FN253]

In decisions subsequent to *Chapman*, however, the Court has not required a "formulaic indication" by state courts in their review of harmless error.²⁵⁴ [FN254] In these cases, at minimum, the Court has required a statement from the lower courts that the harmless error analysis survives federal scrutiny.²⁵⁵ [FN255] Yet in one instance, the Court remanded a case in which the lower court failed to provide a detailed explanation on the record to support its conclusion of harmless error.²⁵⁶ [FN256]

²⁴⁷ *Id.* at 165 (asserting that "if a judge's bias, however manifested, can be separated and substantial evidence of guilt exists, harmless error may apply").

²⁴⁸ 111 S. Ct. 1884 (1991).

²⁴⁹ *Id.* at 1893.

²⁵⁰ *Id.* Justice Souter also assumes that jurors "generally follow the instructions they are given." *Id.* This assumption may not be consistent with empirical research on jury decision making. See Arthur S. Hayes, *Jurors' Grasp of Instructions May Stir Appeal*, WALL ST. J., July 16, 1992, at B1, B5 (citing Zeisel study findings to the contrary).

²⁵¹ For a critique of the application of the harmless error rule, see Francis A. Allen, *A Serendipitous Trek Through the Advance-Sheet Jungle: Criminal Justice in the Courts of Review*, 70 IOWA L. REV. 311, 329-35 (1985) (critiquing application of harmless error rule).

²⁵² To make a judgment of error, a court must make two inquiries: (1) What evidence did the jury consider in reaching its verdict?; and (2) Was the probative force of the noncontested evidence considered by the jury so overwhelming as to leave it beyond a reasonable doubt that the verdict based on that evidence would have been the same in the absence of the contested evidence? *Yates*, 111 S. Ct. at 1893-94.

²⁵³ See *Ogletree*, *supra* note 21, at 159 (reviewing cases) (citing *United States v. Hasting*, 461 U.S. 499 (1983)) (stating that it is "the duty of the court to consider the trial record as a whole"); cf. appendix C (noting Iowa Study findings regarding the reliability and consistency of the on-line ratings).

²⁵⁴ *Sochor v. Florida*, 112 S. Ct. 2114, 2123 (1992); cf. *id.* at 2123 (O'Connor, J., concurring) ("An appellate court's bald assertion that an error of constitutional dimensions was 'harmless' cannot substitute for a principled explanation of how the court reached that conclusion.").

²⁵⁵ *Id.* at 2123 (Souter, J., for the Court) ("[A] plain statement that the judgment survives ... [harmless error analysis] is clearly preferable").

²⁵⁶ See *Clemons v. Mississippi*, 494 U.S. 738, 753-55 (1990) (noting the lower court's "cryptic" one-sentence conclusion to support finding of harmless error).

It is not surprising, therefore, that federal and state appellate courts have not consistently interpreted the circumstances under which error requires reversal of a verdict.²⁵⁷ [FN257] This may be the result of a lack of means for assessing the impact of legal and extralegal factors on trial outcome. It may also be a result of the reality that the assessment of error, whether defined as trial or structural error, is a difficult task for appellate courts. Before error can be held to be harmless, a court must be able to find that it was harmless beyond a reasonable doubt.²⁵⁸ [FN258]

Subsequent tests of the model will explore methods for reviewing courts to more systematically perform error analysis.²⁵⁹ [FN259] Future tests will also examine the ways prosecution and defense counsel may document at trial the nature of the alleged prejudicial behavior or evidence to allow appellate courts to determine from the record whether the error, in whole or in substantial part, contributed to the conviction.²⁶⁰ [FN260] In light of the dearth of empirical information on actual trial error,²⁶¹ [FN261] the preliminary findings from the Iowa Study are encouraging. The findings imply that an appellate court, like the independent on-line review performed by the Iowa observers (or by an "average" juror²⁶² [FN262]), may make reliable evaluations of the relation among legal and extralegal forces, trial outcome, and error.²⁶³ [FN263]

Finally, the model enables systematic study of the relation between alleged error by biased trial judges and principles of disqualification law.²⁶⁴ [FN264] Appellate courts remain reluctant to question the discretion and behavior of trial judges.²⁶⁵ [FN265] Rarely is the behavior of the trial judge alone the subject of review on appeal. Again, this is true in part because there has not been a method to gauge claims that a judge's behavior or conduct has biased the trial against the defendant. It is also likely true because of appellate courts' general reluctance to review motions to disqualify based on trial judges' behavior as proof of bias. Yet, as Professor Leubsdorf has

²⁵⁷ Teitelbaum et al., supra note 215, at 1176-77 & n.60. Discussion here is limited to federal court cases; for a review of related statecourt decisions, see id. at 1178. For a general review of harmless error approaches, see Tom Stacy & Kim Dayton, Rethinking Harmless Constitutional Error, 88 COLUM. L. REV. 79 (1988).

²⁵⁸ See *Chapman v. California*, 386 U.S. 18, 24 (1967) (reviewing when constitutional error requires a reversal); *Arizona v. Fulminante*, 111 S. Ct. 1246, 1265 (1991) (holding that the erroneous admission of a coerced confession is subject to harmless error analysis on appeal); see also Ogletree, supra note 21, at 161-72 (reviewing harmless error rule); Teitelbaum et al., supra note 215, at 1179 & n.67 (providing examples of constitutional and nonconstitutional errors).

²⁵⁹ Teitelbaum et al., supra note 215, at 1177-78 (suggesting that the concept of a "substantial right" is ambiguous and might mean a right to a correct verdict on the evidence as a whole or to a jury verdict that was not influenced by error). For a general test of harmless error in criminal cases, see *Kotteakos v. United States*, 328 U.S. 750 (1946).

²⁶⁰ *The Appearance of Justice*, supra note 1, at 145 (discussing ways that counsel may document record for appellate review of judges' behavior); see also Teitelbaum et al., supra note 215, at 1192 (noting that study is needed of the notion that the right of the accused is not to a perfect trial nor to an irrational verdict, but "to a trial in which improper evidence or argument would not have moved a reasonable jury to convict when otherwise they would have been in doubt").

²⁶¹ Teitelbaum et al., supra note 215, at 1182, 1189 (reporting study using simulated trial scenarios and noting lack of empirical data).

²⁶² See *Harrington v. California*, 395 U.S. 250, 254 (1969) (judging evidence based on probable impact on minds of the "average jury").

²⁶³ Teitelbaum et al., supra note 215, at 1182, 1190-91 (concluding that appellate courts may not always make judgments consistent with juries and, in some circumstances (such as review of a directed verdict), appellate courts must employ their own judgments as to the strength of the evidence against defendants).

²⁶⁴ See Figure 1, supra text accompanying note 40; supra notes 36-40 and accompanying text (discussing relationship between appearance of justice and judicial disqualification or recusal).

²⁶⁵ See *The Appearance of Justice*, supra note 1, at 142-46 (reviewing issues related to appellate court review of judges' behavior).

argued, without the benefit of empirical support, judges' behavior provides "the strongest possible proofs of judicial bias and the proof least subject to evidentiary dispute."²⁶⁶ [FN266]

The model allows reviewing courts to assess alleged judicial bias by enabling them to consider as evidence patterns of behavior, rulings, and decisions.²⁶⁷ [FN267] Likewise, it enables counsel to develop and document the legal and extralegal grounds for a motion alleging judicial bias. The Stanford Study illustrates the usefulness to counsel of employing videotape analyses and the Iowa Study shows the usefulness of employing on-line coding.

In sum, the model allows appellate courts to better calibrate the permissible limits of trial judges' behavior in disqualification cases. Future study of this issue may further public confidence in a core value of our system of justice that requires judges to be fair and independent adjudicators.²⁶⁸ [FN268] The model thus enables review of the values and standards underlying the legal bases for assessing error associated with judicial bias.²⁶⁹ [FN269] Subsequent tests may also assess actual and perceived bias, so as to aid further in the development of disqualification law.²⁷⁰ [FN270] These analyses are warranted, given the findings from the Iowa Study that the appearance of justice alone impacts independently on trial outcomes and sentencing patterns.

C. Research on Sentencing

Prior empirical research has focused primarily on disparity and discrimination in judges' sentencing patterns.²⁷¹ [FN271] Studies conducted before the enactment of the 1987 Federal Sentencing Guidelines show that broad judicial discretion in sentencing often resulted in disparate sentencing patterns.²⁷² [FN272] More recent studies demonstrate that the Federal Sentencing Guidelines have significantly reduced the range of sentences imposed upon different offenders for a similar crimes.²⁷³ [FN273]

Sentencing guidelines allow judges to consider several factors in reaching their decisions. For instance, when the Federal Sentencing Commission established offender categories, they considered, among other things, the individual defendant's age, employment record, and criminal history variables.²⁷⁴ [FN274] But sentencing guidelines are meant to be blind to other factors,

²⁶⁶ Leubsdorf, *supra* note 27, at 275.

²⁶⁷ See *id.* at 275 (suggesting that appellate courts consider a pattern of decisions evidencing bias when ruling on disqualification motions); cf. appendix A (mocking trial proceeding in William Kennedy Smith trial).

²⁶⁸ See Saks & **Blanc**, *supra* note 219, at 830 (reviewing the non-instrumental value of the appearance of justice); Leubsdorf, *supra* note 27, at 278 (pointing out that the appearance standard promotes public confidence in the judiciary); Redish & Marshall, *supra* note 60, at 485-86 (summarizing the various non-instrumental values); cf. *infra* notes 270-90 and accompanying text (discussing results for sentencing).

²⁶⁹ Cf. Leubsdorf, *supra* note 27, at 278 (suggesting that the appearance-of-justice standard alone shifts attention away from objective judicial bias to what appears to be judicial bias to a reasonable observer).

²⁷⁰ *Id.* at 279 (arguing that reliance on the appearance-of-justice standard alone will result in a tendency to disqualify judges in rare cases, such as the case of a judge's verbal or nonverbal expression of bias or public facts that support an inference of bias).

²⁷¹ For a review, see Karle & Sager, *supra* note 50.

²⁷² *Id.* at 395 (summarizing prior empirical research on sentencing and noting that judge background variables were better predictors of the magnitude of the sentence than defendant and crime variables); Ilene H. Nagel, Structuring Sentencing Discretion: The New Federal Sentencing Guidelines, 80 J. CRIM. L. & CRIMINOLOGY 883 (1990) (same).

²⁷³ Karle & Sager, *supra* note 50, at 407.

²⁷⁴ Nagel, *supra* note 272, at 904. These factors are considered only to the extent that they are relevant. *Id.*

such as the defendant's gender, race, and socioeconomic status,²⁷⁵ [FN275] in order to provide uniformity and fairness in the sentencing process.²⁷⁶ [FN276]

The Iowa findings have implications for the study of uniformity and fairness in sentencing. The findings illustrate the impact of legal and extralegal forces on sentencing.²⁷⁷ [FN277] This study, as well as other similar studies, shows that combinations of variables in the model did not accurately predict the magnitude of the sentences.²⁷⁸ [FN278] The Iowa results also comport with findings that sentence disparity in misdemeanor criminal cases is relatively low.²⁷⁹ [FN279] These findings contrast with the historical view that "the sentencing tendency of the judge seems to be fairly well determined before he sits on the bench."²⁸⁰ [FN280]

Future tests of the model will focus on the impact of other background variables on the magnitude of the sentence imposed, controlling for type of case and strength of the evidence. In one study conducted in 1981 by Clancy and his colleagues, greater variance in sentencing disparity was primarily explained by background differences among individual judges rather than by other factors such as offense type or offender characteristics.²⁸¹ [FN281] Clancy also found that judges show relative leniency for younger defendants and are somewhat more harsh for middle-age offenders.²⁸² [FN282] In contrast, the Iowa Study did not find an independent relationship between defendant age and magnitude of the sentence imposed.²⁸³ [FN283] Nor did combinations of the model variables significantly predict the magnitude of the sentence.²⁸⁴ [FN284] Research in more complex federal and state trials is needed to examine the value of the Iowa on-line coding method to the systematic study of fairness and uniformity in the sentencing process.²⁸⁵ [FN285]

Study of the model may also identify particular types of cases in which sentencing disparities are most apparent. For instance, there is a general lack of empirical information, particularly

²⁷⁵ *Id.*

²⁷⁶ *Id.* at 932 (stating that the Guidelines' purpose is to promote uniformity, fairness, certainty, and proportionality).

²⁷⁷ Cf. James L. Gibson, Environmental Constraints on the Behavior of Judges: A Representational Model of Judicial Decision Making, 14 *LAW & SOC'Y REV.* 343, 345 (1980) (focusing on process of decision making and sentencing in the Iowa courts).

²⁷⁸ See *supra* notes 121-73 and accompanying text; see, e.g., Imogen Brown & Roy Hullin, A Study of Sentencing in the Leeds Magistrates Courts: The Treatment of Ethnic Minority and White Offenders, 32 *BRIT. J. CRIMINOLOGY* 41, 52 (1992) (reviewing empirical study in United Kingdom where courts did not find discrimination based on race in sentencing and citing other studies in support of this finding).

²⁷⁹ William Austin & Thomas A. Williams, III, A Survey of Judges' Responses to Simulated Legal Cases: Research Note on Sentencing Disparity, 68 *J. CRIM. L. & CRIMINOLOGY* 306, 309 (1977). Sentencing disparity is related to the type of offense, occurring least in drunken-driving cases and most in drug- cases. *Id.*

²⁸⁰ Frederick J. Gaudet et al., Individual Differences in the Sentencing Tendencies of Judges, 23 *J. CRIM. L. & CRIMINOLOGY* 811, 814 (1933).

²⁸¹ See Kevin Clancy et al., Sentencing Decision Making: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity, 72 *J. CRIM. L. & CRIMINOLOGY* 524, 535, 551-53 (1981) (finding that extralegal factors-such as judges' orientations or views of the criminal justice system- account for 40% of the variance in prison time sentences). The Stanford Study explored the relation among judge background variables, such as age, gender, expectations, and communicative behavior. The findings show that a judge's age and gender relates to his or her communication style with jurors, and this relationship is mediated by other factors in the model. See *The Appearance of Justice*, *supra* note 1, at 124-28.

²⁸² Clancy et al., *supra* note 281, at 539. Defendants over age 60 are given relative leniency. *Id.*

²⁸³ The Iowa Study did find such a relationship with trial outcome. See *supra* notes 136-37 and accompanying text.

²⁸⁴ See *supra* notes 182 & 185 and accompanying text.

²⁸⁵ Cf. Clancy et al., *supra* note 281, at 525 (noting that there is little empirical evidence on sentencing disparity in the federal courts, "and virtually no empirical investigation of the phenomenon on a national scale").

concerning state court sentencing patterns.²⁸⁶ [FN286] A 1980 study that examined sentencing patterns in state courts analyzed the relationship between mode of case disposition (i.e., bench or jury trial) and sentencing for felony crimes, controlling for case seriousness.²⁸⁷ [FN287] The results showed counter-intuitively that harsher sentences were imposed on jury trial defendants than on bench trial defendants. This finding warrants subsequent replication through analyses of other measures. Additionally, analyses will need to delineate the sentencing variable into jail sentences and other nonjail sanctions, such as fines or probation.²⁸⁸ [FN288] These analyses will add to the predictive power of the model by accounting for a broader range of sanctions.²⁸⁹ [FN289]

Finally, the model allows for analyses over time of the factors determining trial outcomes and sentences. For instance, the model allows researchers to calibrate sentencing patterns longitudinally when controlling for factors such as the strength of the evidence or defendant background characteristics.²⁹⁰ [FN290] Additional study with different types of cases is warranted, and future tests of the model will need to delineate the type of case (e.g., felony versus misdemeanor crime) by a measure of charge seriousness.²⁹¹ [FN291]

CONCLUSION

Harry Kalven, Jr. wrote: "The fundamental premise in the idea of impartial judges and rules of law is that certain kinds of decision-making, for example, by judges, can by institutional arrangements and role discipline be made to show less variance and less correlation to personal factors than other kinds of decision-making"²⁹² [FN292] This premise has been tested in the present studies. Some of the findings are encouraging, others require additional study.²⁹³ [FN293]

The research in this Article has explored the legal and extralegal factors related to decision making in bench trials. The findings support the view that many forces, independently and in combination, contribute to decision making and sentencing. The long-term goal is to understand the meaning of terms central to conceptions of justice, such as "trial fairness," "trial error," and

²⁸⁶ See Albert W. Alschuler, *The Trial Judge's Role in Plea Bargaining, Part I*, 76 COLUM. L. REV. 1059, 1085 & n.89 (1976) (noting-before sentencing guidelines were in effect-the limited data on state court sentencing patterns); cf. FEELEY, *supra* note 7, at xvi ("[R]ecurring and continuous phenomenon in lower courts may not occur in higher courts ... [and] there are no published full length comparative studies of lower criminal courts.").

²⁸⁷ Thomas M. Uhlman & N. Darlene Walker, "He Takes Some of My Time; I Take Some of His": An Analysis of Judicial Sentencing Patterns in Jury Cases, 14 LAW & SOC'Y REV. 323 (1980) (reviewing histories of some 30,000 state cases); cf. *supra* note 196 (Iowa results of seriousness of crime as measured by maximum sentence variable).

²⁸⁸ Uhlman & Walker, *supra* note 287, at 327 (suggesting similar analysis).

²⁸⁹ Scales of sentence type and severity have been developed and would enhance the predictive power of the model. *Id.* at 327-28.

²⁹⁰ Cf. Austin & Williams, *supra* note 279, at 309 (discussing a simulated study that assessed sentencing disparity among the same type of judges in the same state jurisdiction and showed that the strength of evidence affects disparity in verdicts and sentences; for instance, where evidence was weak, sentence disparity was higher).

²⁹¹ See *supra* note 196 (discussing results for maximum sentence variable as proxy for case seriousness); see also Uhlman & Walker, *supra* note 287, at 329 (using measure of case seriousness and charge severity). Moreover, analyses may be performed at critical points during the trial. See, e.g., Schulhofer, *supra* note 213, at 1071 (noting that in bench trials closing arguments were a time of active participation by judges).

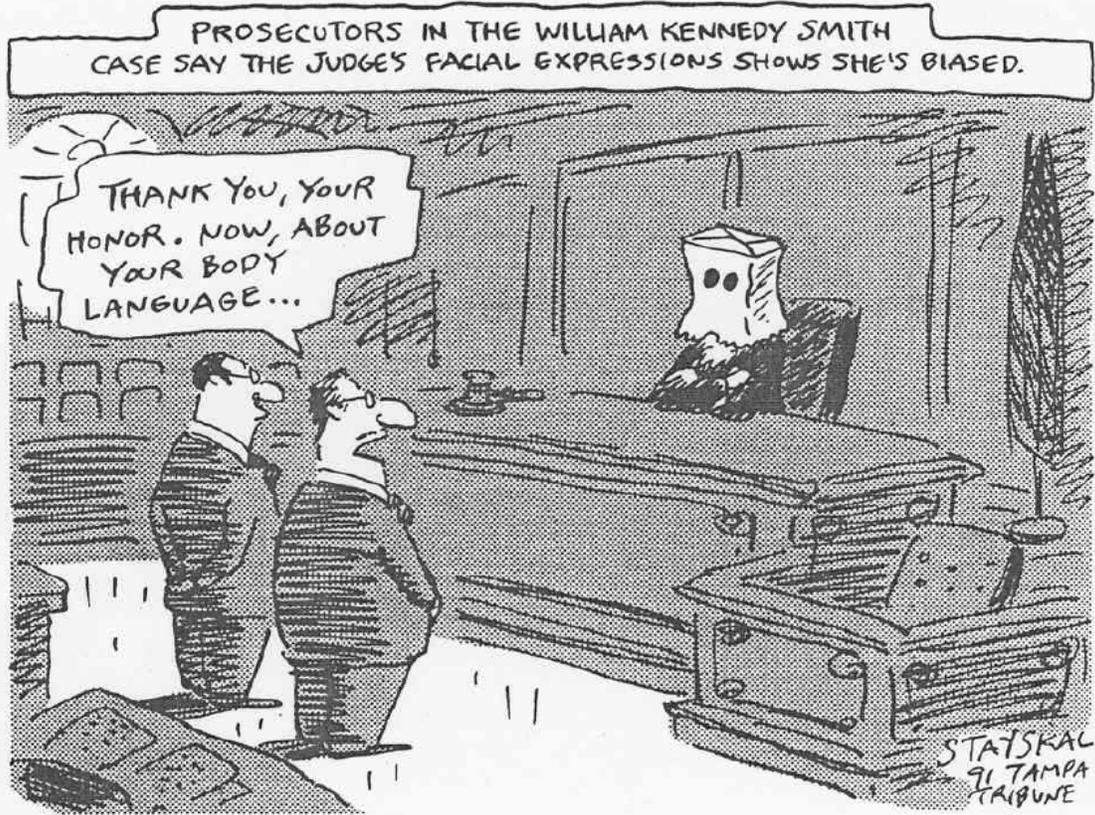
²⁹² Harry Kalven, Jr., *Toward a Science of Impartial Judicial Behavior*, 42 U CIN. L. REV. 591, 594 (1973).

²⁹³ The program of research follows Kalven's views on the relation of law and social science: "Let us empiricize jurisprudence and intellectualize fact finding." *Id.* at 595.

"judicial bias." A more immediate goal is to demonstrate how, in some cases, trial outcomes reflect trial participants' beliefs, attitudes, and biases more than we would like to acknowledge.

APPENDIX A

One View of Nonverbal Behavior in the Courtroom



APPENDIX B

Summary of Data Collection Instruments for Iowa Study of Bench Trials

The following instructions and data collection instruments were provided to all observers. Each observer also received in-depth training on how to use the data collection measures. The data collection measures are summarized below.

1. Trial Information

Name of Coder
Trial Date
Name of Case
Case Number
Name of Defendant
Birthdate of Defendant
Presiding Judge

2. Instructions for Coding This Trial

The following set of instructions are meant to guide you through the process of studying this trial. The information on the cover sheet and on the background variables of trial participants can be obtained through the analysis of the complaint form for the trial, which is located in the Clerk of Misdemeanor's Office. Ask for this form only from this office, never while it is in possession of the presiding magistrate (judge). It is also possible to determine this information from observation of the trial.

The main task of this study is to observe and rate the behaviors of the judge during various parts of the trial. All ratings should be done with the same timing method, with exact coordination of timing between the two raters of the trial. The starting and stopping time for each rating should be the same for each rater; that is, the ratings should be done simultaneously. However, your ratings of the judge's behavior is to be done separate and independently. No consultation between raters should occur on the description of the judge's behavior or any other descriptions about the trial. The ratings you make should be based solely on your own beliefs and perceptions about this trial, as are all other observations.

3. Codebook Table of Contents

Stage of Trial	Actual Page Number in Codebook
I. Background Variables of the Trial Participants	
A. Defendant	1
1. List of Charges in this Trial	1
B. Judge	2
C. Prosecuting Attorney	2

D. Defense Attorney	2
II. Defense's Opening Statement	
A. Micro Behaviors of the Judge	3
B. Global Behaviors of the Judge	4
C. Your Rating of Judge's Expectations of Trial Outcome	5
III. Prosecution's Opening Statement	
A. Micro Behaviors of the Judge	6
B. Global Behaviors of the Judge	7
C. Your Rating of Judge's Expectations of Trial Outcome	8
IV. Direct Examination of the First Witness	
A. Micro Behaviors of the Judge	9
B. Global Behaviors of the Judge	10
C. Your Rating of Judge's Expectations of Trial Outcome	11
D. Record of Objections Made by Counsel	12
V. Cross-Examination of the First Witness	
A. Micro Behaviors of the Judge	13
B. Global Behaviors of the Judge	14
C. Your Rating of Judge's Expectations of Trial Outcome	15
D. Record of Objections Made by Counsel	16
[Index Portion Indicating Same Format for All Witnesses Deleted]	
XII. Prosecution's Closing Statement	
A. Micro Behaviors of the Judge	41
B. Global Behaviors of the Judge	42
C. Your Rating of Judge's Expectations of Trial Outcome	43
XIII. Defense's Closing Statement	
A. Micro Behaviors of the Judge	44
B. Global Behaviors of the Judge	45
C. Your Rating of Judge's Expectations of Trial Outcome	46
XIV. Final Deliberation by the Judge	
A. Micro Behaviors of the Judge	47
B. Global Behaviors of the Judge	48
C. Your Rating of Judge's Expectations of Trial Outcome	49
XV. Verdict for Each Charge Reached by the Jury	50
XVI. Sentence Imposed by the Judge for Each Charge	51
XVII. Final Ratings of the Trial	
A. Rating of Competency for All Trial Participants	52

B. Rating of the Complexity of the Evidence in this Trial	53
C. Rating of the Strength of the Evidence in this Trial	54
D. Description of the Strength of the Evidence	54
E. How Did the Prosecuting Attorney Influence the Trial Outcome?	55
F. How Did the Defense Attorney Influence the Trial Outcome?	56
G. How Did the Judge Influence the Trial Outcome?	57

4. Samples of Codebook Instructions

I. Background Variables of Trial Participants

A. Defendant

1. Sex: Male/Female
2. Age: 18-25, 26-35, 35-45, 46-older
3. Race: _____
4. Socioeconomic Status: High/Medium/Low

List the charges and/or counts in this trial. (List the criminal code sections and the names of charges for each count. This information can be obtained also through review of the complaint form before the trial.)

- a. _____
- b. _____

A. Judge

1. Sex: Male/Female
2. Age: 18-25, 26-35, 35-45, 46-older
3. Race: _____
4. Date appointed to bench: _____

A. Prosecuting Attorney

1. Sex: Male/Female
2. Age: 18-25, 26-35, 35-45, 46-older
3. Race: _____
4. Number of years in practice: _____

A. Defense Attorney or Pro Se

1. Sex: Male/Female
2. Age: 18-25, 26-35, 35-45, 46-older
3. Race: _____
4. Number of years in practice: _____

II. Defense's Opening Statement [Same Format for Defense's Closing Argument and for Prosecutor's Opening Statement and Closing Argument]

A. Micro Behaviors of Judge

Please tally the following micro nonverbal behaviors of the judge for the first five minutes of the defense's opening statement.

Time Start _____ Time Stop _____

1. Amount of eye contact with trial participants (at least 2-3 seconds)

Tallies _____ Total _____

2. Number of smiles

Tallies _____ Total _____

3. Number of significant head nods

Tallies _____ Total _____

4. Number of forward leans

Tallies _____ Total _____

5. Number of significant hand movements

Tallies _____ Total _____

6. Number of significant changes in posture

Tallies _____ Total _____

7. Number of self-touching behaviors

Tallies _____ Total _____

B. Global Behaviors of Judge

Please now rate the following "global" behaviors of the judge for the first five minutes of the defense's opening statement.

not at all judicial 1 2 3 4 5 6 7 8 9 very judicial
(Behavioral examples: professional, dignified, wise, and fair)

not at all directive 1 2 3 4 5 6 7 8 9 very directive
(Behavioral examples: task-oriented, administrative, emphasis on procedure)

not at all warm 1 2 3 4 5 6 7 8 9 very warm
(Behavioral examples: emphatic, supporting, accepting)

not at all confident 1 2 3 4 5 6 7 8 9 very confident
(Behavioral examples: self-assured, interested, emotionally comfortable with other trial participants)

C. Prediction of Judge's Expectations

Based on your reactions to the defense's complete opening statement, what do you believe the judge expects the outcome of the trial to be?

Time _____

	1	2	3	4	5	6	7	8	9
Defendant		Defendant		Not		Defendant		Defendant	
Definitely		Probably		Sure		Probably		Definitely	
Guilty		Guilty				Not Guilty		Not Guilty	

Please give comments and reasons for your rating.

IV. Direct Examination of the First Witness [Same Format for All Witnesses]

A. *Micro Behaviors of Judge*

Please tally the following micro nonverbal behaviors of the judge for the first five minutes of the direct examination of the first witness.

Time start _____ Time stop _____

Witness is testifying for: Defense/Prosecution (circle one)

1. Amount of eye contact with trial participants (at least 2-3 seconds)

Tallies _____ Total _____

2. Number of smiles

Tallies _____ Total _____

3. Number of significant head nods

Tallies _____ Total _____

4. Number of forward leans

Tallies _____ Total _____

5. Number of significant hand movements

Tallies _____ Total _____

6. Number of significant changes in posture

Tallies _____ Total _____

7. Number of self-touching behaviors

Tallies _____ Total _____

B. *Global Behaviors of Judge* Please now rate the following "global" behaviors of the judge for the first five minutes of direct examination of the first witness.

not at all judicial 1 2 3 4 5 6 7 8 9 very judicial
(Behavioral examples: professional, dignified, wise, and fair)

not at all directive 1 2 3 4 5 6 7 8 9 very directive
(Behavioral examples: task-oriented, administrative, emphasis on procedure)

not at all warm 1 2 3 4 5 6 7 8 9 very warm
(Behavioral examples: emphatic, supporting, accepting)

not at all confident 1 2 3 4 5 6 7 8 9 very confident

(Behavioral examples: self-assured, interested, emotionally comfortable with other trial participants)

C. Predictions of Judge's Expectations

Based on your reactions to the direct examination of the first witness, what do you believe the judge expects the outcome of the trial to be?

Time _____

1	2	3	4	5	6	7	8	9
Defendant Definitely Guilty	Defendant Probably Guilty		Not Sure		Defendant Probably Not Guilty		Defendant Definitely Not Guilty	

Please give comments and reasons for your rating.

D. Objections

Record the objections made by counsel during the testimony of this witness in the following format.

Objection # ____.

Objection made by: Defense/Prosecution (Circle One) Description of the objection itself (write it as close as possible to its original wording)

Judge's ruling on this objection: sustained/overruled

XV. Verdict for Each Charge Reached by the Jury

Please record the verdict for each charge or count reached by the Judge at the end of the trial proceedings.

[Coded the Same for Each Count]

1. Charge	Guilty	Not Guilty
a. Count 1	Guilty	Not Guilty
b. Count 2	Guilty	Not Guilty

XVI. Sentence Imposed by the Judge for Each Charge

For each charge or count of which the defendant has been found guilty, state the sentence or penalty imposed by the judge. Also state the maximum possible sentence for each charge. Regular misdemeanor sentence is thirty (30) days in county jail and/or a \$100 fine. There is an exception for a scheduled fine. If the words "scheduled fine" appear on the face of the complaint form, the maximum sentence is always the sentence imposed by the judge.

[Coded the Same for Each Count]

	Sentence	Maximum
1. Charge	_____	_____
a. Count 1	_____	_____
b. Count 2	_____	_____

XVII. Final Ratings of the Trial

A. Rating of Competency for all Trial Participants

Please rate on the following scales the competency of each significant participant in this trial.

1. Competence of the Judge

not at all competent 1 2 3 4 5 6 7 8 9 very competent

2. Competence of the Prosecuting Attorney

not at all competent 1 2 3 4 5 6 7 8 9 very competent

3. Competence of the Defense Attorney

not at all competent 1 2 3 4 5 6 7 8 9 very competent

4. Competence of the Defense's Witnesses

not at all competent 1 2 3 4 5 6 7 8 9 very competent

5. Competence of the Prosecution's Witnesses

not at all competent 1 2 3 4 5 6 7 8 9 very competent

B. Rating of the Complexity of the Evidence

Please rate on the following scale how complex you perceive the evidence in this trial to be. Consider the factual complexity of evidence and the complexity of legal technicalities. Other factors to consider are sophisticated analysis of physical evidence, specialized knowledge, or opinions of significant professionals (e.g., psychologists or medical doctors) on complicated subject matter or legal matters.

not at all complex 1 2 3 4 5 6 7 8 9 very complex

1. Please provide comments on the evidence that was not complex and on evidence that was very complex.

(An example of evidence that is not complex would be direct eyewitness testimony to the crime. Evidence that is complex would be scientific analysis of the crime scene, medical examination of defendant or victim, or complicated legal matters.)

C. Rating of the Strength of the Evidence

Please rate on the following scale the cumulative strength or effect of all the evidence that would lead to a verdict.

strong guilt 1 2 3 4 5 6 7 8 9 weak guilt

D. Description of the Strength of the Evidence

If the evidence was weak toward guilt, please give a description and comments of what was weak about it. If the evidence was strong please give a description and comments of what was strong about it. Please comment on what weak or strong evidence influence a verdict of not guilty or guilty.

E. *Did the Prosecuting Attorney Influence the Trial Outcome?* [Coded the same for defense attorney and judge]

not at all influential 1 2 3 4 5 6 7 8 9 very influential

Please state in a descriptive fashion if the prosecuting attorney influenced the outcome of this trial in either direction, and how he or she did so. (Examples may be: built strong or weak arguments; method of speech was boring or entertaining, interrupted, or fluent; was extremely enthusiastic or interested or was apathetic, etc.)

APPENDIX C

Psychometric Findings

This Appendix highlights the reliability, consistency, and validity of the on-line rating method.

1. Observer Reliability

There is little systematic information available to researchers, courts, and practitioners about the simple reliability and utility (effective reliability) of ratings of on-line courtroom behaviors and processes. An unanswered empirical question is how much of trial behavior needs to be studied or sampled to generalize that behavior to other parts of the trial or to other trials. A related question is the extent to which behavior at critical points in the trial process (e.g., during the testimony of defendant) impacts on assessments of other segments of that trial.²⁹⁴ [FN294] The analyses in this Section begin to develop these issues.

The simple reliability of a single observer is identical to the reliability (i.e., the correlation) between any two observers on a particular variable. The effective or actual reliability is the reliability of the mean of the two observers' observations on a particular variable.²⁹⁵ [FN295] Effective reliability better estimates the agreement among the sample of observers for the measures tested in the model. Simple and effective reliabilities are presented next for several of the variables under study.²⁹⁶ [FN296]

a. Reliability for Background Variables

Predictably, the observers show high agreement in coding the defendants' background variables:

age: r and $er = .99$;
gender: r and $er = 1.0$;
socioeconomic status: $r = .90$, $er = .95$; and
criminal history: $r = .70$, $er = .82$.²⁹⁷ [FN297]

These findings suggest that researchers interested in assessing the objective background variables of trial participants may obtain highly reliable results employing relatively few observers.

b. Reliability for Expectancy Variables

Observers show fairly high agreement in assessing judges' expectations for trial outcomes over the course of entire trials. The reliabilities of the "B" variable during the trial segments are:

²⁹⁴ See *The Appearance of Justice*, supra note 1, at 118 n.97 (calling for future research on the psychometric study of trial processes and behavior).

²⁹⁵ For a review of these concepts, see Robert Rosenthal, *Conducting Judgment Studies*, in *HANDBOOK OF METHODS IN NONVERBAL BEHAVIOR RESEARCH* 287, 292-95 (Klaus R. Scherer & Paul Ekman eds., 1982). The use of effective reliability depends on the assumption that a comparable group of observers would show comparable mean reliability among themselves and with the actual group of ten observers here. *Id.* at 293.

²⁹⁶ Simple reliability is referred to as " r " and effective reliability as " er ."

²⁹⁷ All findings for "A" reliabilities are statistically significant at $p < .0001$.

witness testimony: $r = .51$, $er = .68$;
closing arguments:²⁹⁸ [FN298] $r = .54$, $er = .70$; and
judges' deliberation periods:²⁹⁹ [FN299] $r = .14$, $er = .26$.

The average reliability for the "B" variable is $r = .40$, $er = .57$.³⁰⁰ [FN300] Over the course of entire trials, therefore, researchers should be able to develop reliable estimates of judges' expectations for trial outcomes as perceived by groups of independent observers.

c. Reliability for Communicative Variables

This part presents the simple and effective reliabilities for the communicative variables over the course of entire trials. The purpose is to explore the extent to which on-line observers can reliably assess judges' communicative behavior. The reliabilities of the seven micro behaviors are:

eye contact: $r = .44$, $er = .61$;
head nods: $r = .60$, $er = .75$;
hand movements: $r = .58$, $er = .73$;
postural changes: $r = .20$, $er = .33$;
self-touching: $r = .58$, $er = .73$;
smiles: $r = .05$, $er = .10$; and
forward leans: $r = .21$, $er = .35$.

Judges' smiling behavior is the least reliable micro measure, while eye contact, head nods, hand movements, and self-touching show relatively high reliability. The average reliability for the "C" micro variables is $r = .39$, $er = .56$.³⁰¹ [FN301] These findings add to the literature in other contexts that micro nonverbal variables can be assessed reliably over the course of entire trials.³⁰² [FN302]

The reliabilities of the four global behaviors are:

judicial: $r = .44$, $er = .61$;
directive: $r = .04$, $er = .08$;
warm: $r = .60$, $er = .75$; and
confident: $r = .32$, $er = .48$.

The judicial, warm, and confident styles show fairly high observer reliability, while the directive style is rated much less reliably. The average reliability for the "C" global variables is $r = .35$, $er = .52$.³⁰³ [FN303] The findings for the micro behaviors and global styles suggest that, over the course of entire trials, observers may assess judges' communicative behavior reliably. The findings do not suggest, however, that brief segments of judges' behavior may be naively substituted for assessments made on entire trials. The findings may aid researchers, practitioners, or courts in more effectively developing on-line coding techniques to calibrate the permissible limits of judges' behavior.³⁰⁴ [FN304]

d. Reliability for Evidentiary Variables

²⁹⁸ Sample size for this finding is 11 trials.

²⁹⁹ Sample size for this finding is 18 trials.

³⁰⁰ This average is statistically significant at $p < .001$.

³⁰¹ This average is statistically significant at $p < .001$.

³⁰² See supra note 295.

³⁰³ This average is statistically significant at $p < .001$.

³⁰⁴ See supra notes 238-47 and accompanying text (discussing structural and harmless errors).

The reliabilities for ratings of the strength and complexity of the evidence are as follows:
strength of the evidence:³⁰⁵ [FN305] $r = .64$, $er = .78$; and
complexity of the evidence:³⁰⁶ [FN306] $r = .25$, $er = .40$.

Thus, over the course of these trials, observers are able to agree reliably as to the strength and complexity of the evidence. Because of the central importance of this variable, future researchers might be advised to employ more observers than used here or more detailed analyses of the quality of the evidence (or of evidentiary error) to achieve a higher level of reliability on this variable. The analyses herein focus on the strength of the evidence, which is rated more reliably in this study than the complexity of the evidence.

The finding of relatively high reliability for the strength of the evidence variable is encouraging, given that the ratings are made of actual evidence presented during "live" trials. Prior studies involving simulated and summary evidentiary ratings have suggested that often there is not close agreement regarding ratings of the prejudice associated with various items of evidence.³⁰⁷ [FN307] In one study, members of the bench and community were asked to evaluate the prejudice associated with various evidentiary items in the context of a summarized trial.³⁰⁸ [FN308] The findings show that lawyer and lay participants did not share a common evaluation (e.g., finding low observer reliability) regarding the prejudice associated with various evidentiary items.

The present findings suggest that these on-line observers of courtroom processes agree reliably as to the strength of the evidence. They may also suggest that, to the extent that the independent observers in this study are more like "jurors" than they are counsel or judges, these observers are able to develop a common evaluation (though not necessarily a correct one) of evidentiary strength.³⁰⁹ [FN309]

e. Reliability for Verdict Variable

Predictably, the observers agree perfectly as to the coding of trial outcomes (r and $er = 1.0$, for all 52 trials).

f. Reliability for Sentence Variable

Again, predictably, the observers agree strongly as to the magnitude of the sentence imposed by the judge (r and $er = .99$, for 40 trials).

g. Reliability for Competence/Influence Variables

At the conclusion of the trials, the competence and influence of the judge and prosecution and

³⁰⁵ Sample size for this test is 48 trials. Results are statistically significant at $p < .0001$.

³⁰⁶ Sample size for this test is 50 trials. Results are statistically significant at $p < .0001$.

³⁰⁷ See, e.g., Teitelbaum et al., *supra* note 215, at 1172-76 (finding high disagreement within and across groups of professional and lay observers of evidence).

³⁰⁸ *Id.* at 1156 (noting that the purpose of the study, in part, was to identify items that were high and low in prejudice).

³⁰⁹ But see *id.* at 1163 (implying that jurors may not share common evaluation of evidence in simulated trial study). Of the ten observers in the Iowa Study, eight were law students and two were undergraduate pre-law students.

defense attorneys were rated. The simple and effective reliability of these ratings are:³¹⁰ [FN310]

	<i>Competence</i>		<i>Influence</i>	
	r	er	r	er
Judge ³¹¹	.20	.33	.18	.31
Prosecution Counsel ³¹²	.06	.11	.37	.54
Defense Counsel ³¹³	.84	.91	.70	.82
Prosecution Witness ³¹⁴	.37	.54	----	----
Defense Witness ³¹⁵	.44	.61	----	----

The central conclusion that can be drawn from these findings is that there is a good deal of variability in the ratings of the ability of trial participants. For instance, there is only moderate agreement as to judge competence and influence. Subsequent researchers may consider employing more observers in the collection of additional information regarding the subjective and actual competencies of trial participants.³¹⁶ [FN316]

2. Consistency of the Ratings During the Trial

The analyses in this section explore the consistency of the ratings of the "B" and "C" variables over the segments of the trial—for example, across witness testimony, closing arguments, and deliberation period.³¹⁷ [FN317] To assess the consistency of these variables, intraclass correlations averaged over the ratings of the observers are computed. As in Section 1 of this Appendix, the effective reliability of the mean of the three segments of the trials is also presented.

a. Consistency for Expectancy Variables

The simple reliability of any single segment for the "B" variable is .83. The effective reliability of the mean of the segments of the trials assessed is .94. These findings suggest that the ratings of judges' expectations are assessed consistently across the segments of these trials.

b. Consistency for Communicative Variables

The simple and effective reliabilities illustrate the degree to which judges' behavior may be assessed consistently across the three trial segments (e.g., witnesses, closing statement, and deliberation). The reliabilities of the seven micro behaviors are:

³¹⁰ Blank slots in the table indicate that sufficient data were not available.

³¹¹ Sample size is 49 trials.

³¹² Sample size is 49 trials.

³¹³ Sample size is 8 trials, therefore these results are presented for exploratory purposes and must be viewed with caution. See supra note 112 and accompanying text.

³¹⁴ Sample size is 42 trials. Data for influence measure is not available.

³¹⁵ Sample size is 30 trials. Data for influence measure is not available.

³¹⁶ For systematic technique for assessing individual and group behaviors, see ROBERT F. BALES & STEPHEN P. COHEN, SYMLOG: A SYSTEM FOR THE MULTIPLE LEVEL OBSERVATION OF GROUPS (1979).

³¹⁷ For a review of similar analyses, see The Appearance of Justice, supra note 1, at 118 n.97. The methodological question of interest to researchers, practitioners, and courts is whether "B" and "C" variables can be assessed consistently (over time) from the five minute segments of these trials.

eye contact: $r = -.24$, $er = na$,³¹⁸ [FN318]
head nods: $r = .48$, $er = .73$;
hand movements: $r = .59$, $er = .81$;
postural changes: $r = .27$, $er = .53$;
self-touching: $r = .22$, $er = .46$;
smiles: $r = .00$, $er = .00$,³¹⁹ [FN319] and
forward leans: $r = .23$, $er = .47$.

Hand movements and head nods show relatively high consistency, yet smiling behavior and eye contact is not assessed consistently. The average reliability for the micro behaviors is $r = .22$, $er = .46$.³²⁰ [FN320] These findings add to the literature in other contexts showing that many micro nonverbal variables can be assessed consistently over various trial segments.

The reliabilities of the four global styles are:

judicial: $r = .91$, $er = .97$;
directive: $r = .84$, $er = .94$;
warm: $r = .94$, $er = .98$; and
confident: $r = .94$, $er = .98$.

All four global styles are highly consistent across the trial segments.³²¹ [FN321] The average reliability for the global variables is $r = .91$, $er = .97$.³²² [FN322] These findings again suggest that judges' global styles can be assessed consistently over the trial segments.³²³ [FN323]

1. Summary

The purpose of this Appendix was to further the analysis of the reliability and consistency of the variables in the model of courtroom behavior. Consistent with earlier suggestions,³²⁴ [FN324] brief segments of trials may not be naively substituted as effective predictors of other portions of the trial process. Nevertheless, the findings suggest that most "B" and "C" variables, in particular judges' global behavior, may be assessed reliably and consistently by on-line ratings of trials. It is encouraging also that the consistency of the judges' behavioral styles could be assessed reliably from only the first five minutes of the various trial segments.

Further study is warranted to understand the implications of the degree of reliability and

³¹⁸ The er is not applicable ("na") to a negative r .

³¹⁹ The finding of zero reliability for smiles is unpredicted and requires further fine-grain analysis. Subsequent analyses of this micro variable must be viewed with caution because the observers vary significantly in their assessment of smiling behavior.

³²⁰ This average is statistically significant at $p < .01$.

³²¹ The overall correlation coefficient between "judicial" and "confident" is .89. Collinearity or multicollinearity occurs when an independent variable (regressor) is a linear combination of other independent variables (e.g., two global variables). In such a case, the regression coefficients (e.g., Tables 1-4 supra) can be unstable. See D.A. BELSLEY ET AL., REGRESSION DIAGNOSTICS (1980).

To assess the collinearity of the independent measures in the model, SAS Regression Collinearity Diagnostics were employed. These analyses explore high correlations between independent measures. The results show that only the pairing of judicial and confident global variables have a relatively high coefficient for the same principal component. For theoretical purposes, and consistent with the Stanford Study, it was decided in the Iowa Study to retain the four global behaviors as independent variables in the model.

³²² This average is statistically significant at $p < .001$.

³²³ See supra notes 58-64 and accompanying text (discussing "quantifying" judges' error).

³²⁴ The Measure of the Judge, supra note 1, at 666 n.63 (noting that more research is needed before any conclusive statement may be made about the generalizability of videotape analyses to other parts of trials).

consistency in judges' styles in relating to juries and other trial participants. The model allows for an assessment of the relation between consistency in judges' behavior and perceived or actual bias or trial error by the judge. It may allow also for an examination of the relation among judges' or lawyers' consistency (e.g., effectiveness), trial outcomes, and sentencing patterns.