

4-2009

# The "Hidden Judiciary": An Empirical Examination of Executive Branch Justice

Chris Guthrie

*Vanderbilt University Law School*, [Chris.guthrie@vanderbilt.edu](mailto:Chris.guthrie@vanderbilt.edu)

Jeffrey J. Rachlinski

*Cornell Law School*, [jjr7@cornell.edu](mailto:jjr7@cornell.edu)

Andrew J. Wistrich

*U.S. District Court, Central District of California*

Follow this and additional works at: <http://scholarship.law.cornell.edu/facpub>

 Part of the [Administrative Law Commons](#), [Judges Commons](#), and the [President/Executive Department Commons](#)

---

## Recommended Citation

Guthrie, Chris; Rachlinski, Jeffrey J.; and Wistrich, Andrew J., "The "Hidden Judiciary": An Empirical Examination of Executive Branch Justice" (2009). *Cornell Law Faculty Publications*. Paper 1368.

<http://scholarship.law.cornell.edu/facpub/1368>

This Article is brought to you for free and open access by the Faculty Scholarship at Scholarship@Cornell Law: A Digital Repository. It has been accepted for inclusion in Cornell Law Faculty Publications by an authorized administrator of Scholarship@Cornell Law: A Digital Repository. For more information, please contact [jmp8@cornell.edu](mailto:jmp8@cornell.edu).

# THE “HIDDEN JUDICIARY”: AN EMPIRICAL EXAMINATION OF EXECUTIVE BRANCH JUSTICE

CHRIS GUTHRIE†

JEFFREY J. RACHLINSKI††

ANDREW J. WISTRICH†††

## ABSTRACT

*Administrative law judges attract little scholarly attention, yet they decide a large fraction of all civil disputes. In this Article, we demonstrate that these executive branch judges, like their counterparts in the judicial branch, tend to make predominantly intuitive rather than predominantly deliberative decisions. This finding sheds new light on executive branch justice by suggesting that judicial intuition, not judicial independence, is the most significant challenge facing these important judicial officers.*

## TABLE OF CONTENTS

Introduction .....	1478
I. Intuitive-Override Model of Judging .....	1480
II. Intuition and Deliberation in Administrative Law Judges.....	1482
A. Expertise .....	1483
B. Accountability .....	1486
C. Feedback .....	1489
D. Summary .....	1491

---

Copyright © 2009 by Chris Guthrie, Jeffrey J. Rachlinski, and Andrew J. Wistrich.

† Chris Guthrie is a Professor of Law at Vanderbilt Law School.

†† Jeffrey J. Rachlinski is a Professor of Law at Cornell Law School.

††† Andrew J. Wistrich is a Magistrate Judge of the United States District Court for the Central District of California.

We thank Professor Lawrence Baum, Judge Gerard Lynch, Judge Barbara Rothstein, and other participants in this Symposium for helpful feedback on this Article.

III. Studying the Administrative Law Judge.....	1491
A. Methods.....	1492
B. Results: Cognitive Reflection.....	1495
1. Cognitive Reflection Test.....	1495
2. CRT and ALJs.....	1498
C. Results: Judicial Decision Making.....	1501
1. Anchoring.....	1501
2. Framing and Fairness.....	1506
3. Conjunction.....	1509
4. Hindsight/Outcome Bias.....	1512
5. Disregarding.....	1516
6. Egocentric Bias.....	1518
D. Summary.....	1520
Conclusion.....	1522
Appendix A: Materials Used at the New York City Conference..	1524
Appendix B: Materials Used at the National Conference.....	1527

## INTRODUCTION

Administrative law judges (ALJs) toil in the shadows of the civil justice system. They work for the executive branches of state and federal governments, usually embedded in specialized agencies. Located outside the courtrooms in which generalist judges preside, they comprise a “hidden judiciary.”<sup>1</sup>

To citizens, however, the *14,100* ALJs<sup>2</sup> are anything but hidden. They “adjudicate massive numbers of individual disputes, far exceeding the number resolved by courts.”<sup>3</sup> They handle matters in many areas of concern to citizens and society, including “disability, retirement, and other income security entitlements; consumer, workplace, and environmental safety; labor relations and civil rights violations; and regulatory programs in industry, commerce,

1. See Thomas C. Mans, *Selecting the ‘Hidden Judiciary’: How the Merit Process Works in Choosing Administrative Law Judges (Part I)*, 63 JUDICATURE 60, 60 (1979) (coining the term “hidden judiciary”). But see Ronnie A. Yoder, *The Role of the Administrative Law Judge*, 22 J. NAT’L ASS’N ADMIN. L. JUDGES 321, 323 (2002) (“We [administrative law judges] used to be referred to as the hidden judiciary; but you do not see that phraseology much any more.”).

2. U.S. DEP’T OF LABOR, OCCUPATIONAL EMPLOYMENT AND WAGES, 2007, at 9 tbl.1 (Bureau of Labor Statistics, News Release No. 08-0620, rev. ed. 2009), available at <http://www.bls.gov/news.release/pdf/ocwage.pdf> (reporting that 14,100 “administrative law judges, adjudicators, and hearing officers” were employed in the United States in 2007).

3. Charles H. Koch, Jr., *Policymaking by the Administrative Judiciary*, 56 ALA. L. REV. 693, 693 (2005).

communications, banking, and transportation.”<sup>4</sup> At the federal level, ALJs conduct at least nine times as many trials as federal judges.<sup>5</sup> In short, ALJs are the “face of justice” for most American citizens.<sup>6</sup>

The prevalence of administrative judging exposes a great irony concerning ongoing debates about judicial specialization. Scholars, pundits, and politicians periodically call for greater specialization among judges,<sup>7</sup> typically motivated by concerns that generalist judges simply cannot master the many complex areas of law and fact involved in modern litigation. Although the most visible courts are composed of generalist judges, a surprisingly large percentage of disputes are adjudicated by specialist judges. The quiet delegation of judicial authority to administrative tribunals is a long-term trend that has arisen more out of necessity than out of a careful assessment of the benefits and costs of judicial specialization. The wisdom of this trend has received little serious consideration or empirical study. Using well-established psychological research methods, we seek to begin to fill that gap in this Article.

Based on previous research involving generalist judges—federal district judges, state court judges, and federal magistrate judges—we have developed a model of judicial decisionmaking that explains how even well-qualified, experienced, and well-intentioned judges can make erroneous decisions. As we discuss in Part I, we have found that generalist judges appear to rely too heavily on intuition, rather than deliberation, when making judgments.<sup>8</sup> We raise the theoretical

---

4. Daniel L. Skoler, *The Administrative Law Judiciary: Change, Challenge, and Choices*, 462 ANNALS AM. ACAD. POL. & SOC. SCI. 34, 36 (1982).

5. See Judith Resnik, *Migrating, Morphing, and Vanishing: The Empirical and Normative Puzzles of Declining Trial Rates in Courts*, 1 J. EMPIRICAL LEGAL STUD. 783, 799 (2004) (reporting that while federal judges conducted a total of about 85,000 trials (defined as “contested hearings at which evidence is presented”) during 2001, just four federal administrative agencies (the SSA, the INS, the Board of Veterans Appeals, and the EEOC) conducted more than 720,000 trials that year); Paul R. Verkuil, *Reflections upon the Federal Administrative Judiciary*, 39 UCLA L. REV. 1341, 1343 (1992) (“ALJs probably decide more ‘cases’ each year than do their federal judicial counterparts.”).

6. Yoder, *supra* note 1, at 323.

7. See, e.g., Jeffrey J. Rachlinski, Chris Guthrie & Andrew J. Wistrich, *Inside the Bankruptcy Judge’s Mind*, 86 B.U. L. REV. 1227, 1228 (2006) (reviewing the demand for and the trends toward judicial specialization).

8. See generally Chris Guthrie, Jeffrey J. Rachlinski & Andrew J. Wistrich, *Blinking on the Bench: How Judges Decide Cases*, 93 CORNELL L. REV. 1 (2007) [hereinafter Guthrie et al., *Blinking*] (proposing an intuitive-override model of decisionmaking and presenting relevant experimental evidence); Chris Guthrie, Jeffrey J. Rachlinski & Andrew J. Wistrich, *Inside the Judicial Mind*, 86 CORNELL L. REV. 777 (2001) [hereinafter Guthrie et al., *Judicial Mind*] (reporting experimental evidence showing that judges are susceptible to misleading heuristics

possibility in Part II that administrative law judges—who tend to have greater subject-matter expertise and face more frequent decisionmaking oversight—might make better decisions than generalist judges. Part III then describes our empirical study of ALJ decisionmaking. In it, we attempt to sort out whether ALJs do in fact make more deliberative decisions than their generalist judge counterparts. We base our results on two different empirical investigations. First, as Part III.B explains, we report the results of a study designed to assess whether ALJs demonstrate a propensity to engage in impressionistic and intuitive problem solving or deliberative and reflective problem solving when faced with generic decision problems. Second, as Part III.C explains, we report the results of several experiments designed to assess whether ALJs demonstrate susceptibility or resistance to common “heuristics and biases”<sup>9</sup> that have been shown to influence judicial decisionmaking. With regard to both, we find that ALJs, just like generalist judges, make predominantly intuitive rather than predominantly deliberative decisions and are vulnerable to the same types of decisionmaking errors.

We conclude by arguing that our findings shed new light on executive branch justice. Among ALJs, executive branch agencies, and the scholars who study them, judicial independence has dominated discussion. We argue, instead, that the primary issue of concern is *not judicial independence* but *judicial intuition*.

## I. INTUITIVE-OVERRIDE MODEL OF JUDGING

Building on research from psychology as well as our own previous research on judicial decisionmaking, we contend that generalist judges approach legal problems using two distinct systems of judgment.<sup>10</sup> On the one hand, judges, like other human beings, use

---

and biases when making judgments); Andrew J. Wistrich, Chris Guthrie & Jeffrey J. Rachlinski, *Can Judges Ignore Inadmissible Information? The Difficulty of Deliberately Disregarding*, 153 U. PA. L. REV. 1251 (2005) (reporting experimental evidence showing that judges have difficulty deliberately disregarding relevant but inadmissible evidence when making merits decisions).

9. Amos Tversky & Daniel Kahneman, *Judgment Under Uncertainty: Heuristics and Biases*, 185 SCIENCE 1124, 1124 (1974) (introducing the concept of heuristics and biases).

10. See Guthrie et al., *Blinking*, *supra* note 8, at 6; Keith E. Stanovich & Richard F. West, *Individual Differences in Reasoning: Implications for the Rationality Debate?*, in HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT 421, 436 (Thomas Gilovich, Dale Griffin & Daniel Kahneman eds., 2002) (“Although there are several two-process theories of

an intuitive system of judgment, which is made up of so-called “System 1” processes.<sup>11</sup> System 1 processes are “automatic, heuristic-based, and relatively undemanding of computational capacity.”<sup>12</sup> In short, they are “spontaneous, intuitive, effortless, and fast.”<sup>13</sup> When assessing legal problems, judges quickly develop an intuition or “hunch” about the right outcome.<sup>14</sup>

On the other hand, judges, like other human beings, also use a deliberative system of judgment, which is made up of so-called “System 2” processes.<sup>15</sup> System 2 processes require “effort, motivation, concentration, and the execution of learned rules.”<sup>16</sup> In short, these processes are “deliberate, rule-governed, effortful, and slow.”<sup>17</sup> This means that when judges confront legal problems, they are also able to analyze them in a more measured, deliberative, and rule-based way.

Our intuitive-override model of judging asserts that these two systems interact in a fairly predictable way. Relying on a generic model of decisionmaking proposed by Professors Daniel Kahneman and Shane Frederick,<sup>18</sup> we contend that judges initially make intuitive or System 1 judgments—which are effortless, fast, and often accurate—that they *might* override with deliberative or System 2 processes—which are more time and labor intensive. As Kahneman and Frederick explain their model, “System 1 quickly proposes intuitive answers to judgment problems as they arise, and System 2 monitors the quality of these proposals, which it may endorse, correct, or override. The judgments that are eventually expressed are

---

reasoning that differ somewhat in their details, all agree on the general features of the two systems which, for simplicity, we label System 1 and System 2.”)

11. See Stanovich & West, *supra* note 10, at 436.

12. *Id.*

13. Daniel Kahneman & Shane Frederick, *Representativeness Revisited: Attribute Substitution in Intuitive Judgment*, in HEURISTICS AND BIASES: THE PSYCHOLOGY OF INTUITIVE JUDGMENT, *supra* note 10, at 49, 49.

14. See Joseph C. Hutcheson, Jr., *The Judgment Intuitive: The Function of the “Hunch” in Judicial Decision*, 14 CORNELL L.Q. 274, 278 (1929).

15. See Stanovich & West, *supra* note 10, at 436 (“System 2 conjoins the various characteristics associated with controlled processing.”).

16. Shane Frederick, *Cognitive Reflection and Decision Making*, 19 J. ECON. PERSP. 25, 26 (2005).

17. Kahneman & Frederick, *supra* note 13, at 49.

18. *Id.* at 51.

called *intuitive* if they retain the hypothesized initial proposal without much modification.<sup>19</sup>

The intuitive-override model means that judges often rely on simple heuristics, attend closely to various informational cues that can be misleading, and can be influenced by untoward psychological phenomena in their decisionmaking.<sup>20</sup> But judges can sometimes override these misleading intuitive responses with a more deductive, deliberative decisionmaking approach.<sup>21</sup> This combination of intuition and deliberative override makes judges efficient decisionmakers, but we have found evidence suggesting that generalist judges may over-rely on their intuitive processes and under-use their deliberative faculties.

We recognize, of course, that the intuitive and deliberative systems can, and do, interact in a variety of ways.<sup>22</sup> We also recognize that the brain is not neatly divided into two systems—it contains many areas that seem to perform discrete functions that are all interconnected.<sup>23</sup> The notion of two systems is perhaps better viewed as a metaphor, or organizing principle, than a hard-wired reality in the brain. In the main, however, we think that our proposed model captures judges' basic approach to decisionmaking and offers a more realistic, though still tractable, account than those associated with formalism (that is, purely deductive decisionmaking) and realism (that is, intuitive rationalization).

## II. INTUITION AND DELIBERATION IN ADMINISTRATIVE LAW JUDGES

Does the same model of judging apply to ALJs? We suspect that it does. ALJs possess educational and professional credentials comparable to those of generalist judges,<sup>24</sup> and in the courtroom, they

---

19. *Id.*

20. Guthrie et al., *Blinking*, *supra* note 8, at 33.

21. *Id.* at 3.

22. *Id.* at 8–9 (identifying various ways in which intuition and deliberation can interact in judgment and decisionmaking).

23. See Mark Lubell et al., *Institutional Design Capitalizing on the Intuitive Nature of Decision Making*, in *BETTER THAN CONSCIOUS: DECISION MAKING, THE HUMAN MIND, AND IMPLICATIONS FOR INSTITUTIONS* 413, 414 (Christoph Engel & Wolf Singer eds., 2008) (“[T]he distinction between conscious and unconscious is a somewhat artificial construct. So, too, is the supposed ‘system 1’ versus ‘system 2’ distinction.” (citation omitted)).

24. Most ALJs, like most generalist judges, are lawyers, *see, e.g.*, Skoler, *supra* note 4, at 36 (“[Federal] ALJs must be lawyers.”), but some state ALJs are not, *see, e.g.*, Yoder, *supra* note 1,

perform the same functions.<sup>25</sup> That is, they “hear cases, find the facts, and apply the law.”<sup>26</sup> But even if the ALJs rely upon the same combination of intuition and deliberation, several features of the unique role they play in the legal system might lead them to rely more heavily than ordinary judges on deliberation. First, because ALJs tend to specialize, they might develop the *expertise* to discern when it is important to override their intuition with deliberation. Second, because ALJs face greater scrutiny of their decisions than do their generalist counterparts, they might feel more accountable, and this *accountability* might lead to more deliberative decisionmaking. Third, this heightened scrutiny might give ALJs greater access to *feedback*, which might also facilitate more deliberative decisionmaking.<sup>27</sup>

### A. *Expertise*

The most striking difference between most ALJs and most generalist judges is that ALJs almost always *specialize*.<sup>28</sup> In the federal government and many state governments, ALJs work for a particular

---

at 325 (“We come in all shapes and sizes[:] lawyers and non-lawyers, federal and state, central panels and agency employees.”).

25. *Butz v. Economou*, 438 U.S. 478, 513 (1978) (“There can be little doubt that the role of the modern federal hearing examiner or administrative law judge within this framework is ‘functionally comparable’ to that of a judge. His powers are often, if not generally, comparable to those of a trial judge: He may issue subpoenas, rule on proffers of evidence, regulate the course of the hearing, and make or recommend decisions.”); James E. Moliterno, *The Administrative Judiciary’s Independence Myth*, 41 WAKE FOREST L. REV. 1191, 1196–97 (2006) (observing that ALJs “possess, of course, the fundamental core of the judicial definition, which, while nowhere given authoritatively, is by wide approval known to consist in the impartial adjudication of cases”).

26. Michael Asimow, *The Administrative Judiciary: ALJ’s in Historical Perspective*, 20 J. NAT’L ASS’N ADMIN. L. JUDGES 157, 157 (2000); see also R. Terrence Harders, *Striking a Balance: Administrative Law Judge Independence and Accountability*, J. NAT’L ASS’N ADMIN. L. JUDGES, Spring 1999, at 1, 9 (“[T]hose who come before ALJs recognize that the deciding of disputes, the prescribing of duties, and the recognition of entitlement affects them in much the same way that litigation plays out in courts of law.”).

27. Although we focus on potential advantages that ALJs possess, we acknowledge that ALJs might face potential disadvantages as well. For example, ALJs frequently face crowded dockets and often have fewer resources, such as judicial clerks, than many of their generalist counterparts. This might make them more, not less, prone to making snap, routinized, bureaucratic judgments. See Koch, *supra* note 3, at 693 (“Administrative agencies adjudicate massive numbers of individual disputes, far exceeding the number resolved by courts.”).

28. We recognize that a meaningful fraction of judicial branch judges serve on specialized courts, but most judicial branch judges are generalists. See generally CENT. EUROPEAN & EURASIAN LAW INITIATIVE, AM. BAR ASS’N, SPECIALIZED COURTS: A CONCEPT PAPER (1996) (on file with the *Duke Law Journal*) (providing background on the composition of the U.S. specialized courts).



agency, like the Social Security Administration or the Department of Labor.<sup>29</sup> Before ascending to the bench, they often have relevant expertise and practice experience.<sup>30</sup> Once they become ALJs, their expertise in the subject matter of the agency deepens because most of them adjudicate disputes involving only the agency for which they

---

29. See L. Harold Levinson, *The Status of the Administrative Judge*, 38 AM. J. COMP. LAW 523, 535 (1990) (“Most ALJs specialize in adjudicating cases for only one agency . . .”). This is uniformly true in the federal government because there all federal ALJs are employed by a particular agency. In the states, the employment arrangements of ALJs are more complicated and less amenable to “scholarly generalizations.” Jim Rossi, *Overcoming Parochialism: State Administrative Procedure and Institutional Design*, 53 ADMIN. L. REV. 551, 553 (2001); see also Skoler, *supra* note 4, at 35 (“Writing about administrative law judges at both federal and state levels is a tricky business and therefore rarely done.”). Historically, most state ALJs, like their federal counterparts, worked for specific agencies. See, e.g., James F. Flanagan, *Redefining the Role of the State Administrative Law Judge: Central Panels and Their Impact on State ALJ Authority and Standards of Agency Review*, 54 ADMIN. L. REV. 1355, 1357 (2002) [hereinafter Flanagan, *Redefining the Role*] (observing that California created the first central panel in 1945 and that seven states had central panels by 1983). Beginning in the 1980s and 1990s, many more states—now about thirty of them—moved to adopt, in various forms, “central panels” or independent ALJ agencies. See, e.g., James F. Flanagan, *An Update on Developments in Central Panels and ALJ Final Order Authority*, 38 IND. L. REV. 401, 403–04 (2005) [hereinafter Flanagan, *Update on Developments*] (“Twenty-five states, and three major cities, have established central panels thus far.”); Rossi, *supra*, at 568 (“In more than thirty states, to one degree or another administrative law judges (ALJs) are housed in a central panel . . .”); Nat’l Ass’n of Admin. Law Judiciary, *Non-Central Panel States*, <http://www.naalj.org/nonpanel.html> (last visited Mar. 23, 2009) (identifying 29 states with central panels). Professor James Flanagan describes these central panel arrangements as follows:

A central panel of ALJs is a cadre of professional adjudicators who are administratively independent of the agencies whose cases they hear, and thus, they are removed from agency influence. The central panels are organized in several ways. In some states, the central panel is an independent agency within the executive branch. In others, the central panel is part of another agency for administrative support, but independent for all other purposes. A third variation puts the ALJs in a separate organization, but assigns each ALJ to a particular agency based upon expertise in the subject matter.

Flanagan, *Redefining the Role*, *supra*, at 1356; see also William R. Andersen, *Judicial Review of State Administrative Action—Designing the Statutory Framework*, 44 ADMIN. L. REV. 523, 554–57 (1992) (proposing three models of administrative adjudication). *But see* Koch, *supra* note 3, at 733 (“[P]anel judges, since they serve many agencies, are generalists and thus do not provide the expertise and experience inherent in the traditional scheme.”). At present, then, all federal ALJs and nearly half of the state ALJs work within one agency, and among those state ALJs who are employed by central panels, many of them also specialize in the law and policy of a particular agency.

30. See Flanagan, *Redefining the Role*, *supra* note 29, at 1406 (“States often require ALJs to have substantive knowledge or experience in the subject matter . . .”); see also, e.g., N.J. STAT. ANN. § 52:14F-12(a) (West 2001) (requiring that judges hired into the environmental unit have “special expertise” in environmental law); TEX. GOV’T CODE ANN. § 2003.049(d) (Vernon 2000) (requiring ALJs to have experience in utility law).

work, meaning that they hear similar matters over and over again.<sup>31</sup> In contrast to federal district judges—who at any given time might be hearing diversity cases involving various states’ substantive laws, multiple criminal matters, and disputes filed under a wide variety of federal statutes—most ALJs repeatedly hear the same kinds of cases.

We do not want to overstate the difference between ALJs and generalist judges.<sup>32</sup> As our own data show,<sup>33</sup> some of the ALJs hear many different kinds of cases, particularly those who serve on “central panels.” These ALJs may differ little from the trial judges who sit in their states’ court of general jurisdiction. Conversely, in some jurisdictions, state trial judges specialize, devoting themselves to particular areas of law such as juvenile or family law. These state court judges may be no different from many of the ALJs who specialize in narrow areas of law. Nonetheless, specialization is the norm among ALJs but the exception among judicial branch trial judges.

Because ALJs typically develop specialized expertise that their counterparts in the generalist judiciary do not possess, they might be better decisionmakers. Specialization does not always improve decisionmaking, however. Some research shows that greater expertise facilitates better judgment,<sup>34</sup> but other research shows that generalists are just as competent as specialists.<sup>35</sup> In one previous study, we found

---

31. This is so not only for the obvious reason that ALJs are repeatedly exposed to similar matters, but also because the agencies that employ them often provide the ALJs with important guidance to keep them abreast of relevant legal and policy developments. Flanagan, *Redefining the Role*, *supra* note 29, at 1406; *see also, e.g.*, TEX. GOV’T CODE ANN. § 2001.058(c) (providing a specific example requiring the agency to provide a statement on applicable rules and policies).

32. *See* Edward Cheng, *The Myth of the Generalist Judge*, 61 STAN. L. REV. 519, 540 (2008) (presenting evidence that generalist appellate judges tend to write opinions in various specialized areas).

33. *See infra* Part III.A.

34. *See, e.g.*, Chris Guthrie, *Panacea or Pandora’s Box?: The Costs of Options in Negotiation*, 88 IOWA L. REV. 601, 641–42 (2003) (finding that lawyers were less susceptible than nonlawyers to the pernicious effects of contrast or asymmetric dominance); Chris Guthrie & Jeffrey J. Rachlinski, *Insurers, Illusions of Judgment & Litigation*, 59 VAND. L. REV. 2017, 2047 (2006) (“This paper reports the results of several litigation problems—two anchoring problems, three framing problems, and one self-serving bias problem—involving nearly two hundred participants from the insurance industry. The results reported in the paper suggest that these experts, relative to others who have been studied, make decisions that more closely approximate rational choice.”); Russell Korobkin & Chris Guthrie, *Psychology, Economics, and Settlement: A New Look at the Role of the Lawyer*, 76 TEX. L. REV. 77, 99–101 (1997) (finding that lawyers are less susceptible than nonlawyers to framing effects).

35. SCOTT PLOUS, *THE PSYCHOLOGY OF JUDGMENT AND DECISION MAKING* 258 (1993) (“[S]everal studies have found that experts display either roughly the same biases as college

that federal bankruptcy judges appear as susceptible to common errors in judgment as their generalist counterparts,<sup>36</sup> suggesting that specialization “does not lead inexorably to improved decision making.”<sup>37</sup> Nonetheless, typical ALJs, even as compared to bankruptcy judges, tend to develop unusually high levels of specialized expertise and may therefore have much more familiarity with the matters they encounter in court.

### B. Accountability

Even if specialization does not necessarily lead to better decisionmaking, ALJs might make more rational and deliberative decisions than their generalist-judge counterparts because they are more *accountable* for their decisions.<sup>38</sup> In contrast to generalist judges, ALJs often do not make “final” decisions; instead, they issue recommended decisions that are reviewed by their agency heads.<sup>39</sup> As

---

students or the same biases at somewhat reduced levels.”); *see also, e.g.*, John C. Anderson, D. Jordan Lowe & Philip M.J. Reckers, *Evaluation of Auditor Decisions: Hindsight Bias Effects and the Expectation Gap*, 14 J. ECON. PSYCHOL. 711, 725–31 (1993) (indicating that auditors, like nonexperts, are influenced by the hindsight bias); Hal R. Arkes et al., *Eliminating the Hindsight Bias*, 73 J. APPLIED PSYCHOL. 305, 307 (1988) (demonstrating that psychologists, like nonexperts, are prone to the hindsight bias); Linda Babcock et al., *Forming Beliefs About Adjudicated Outcomes: Perceptions of Risk and Reservation Values*, 15 INT'L REV. L. & ECON. 289, 296–97 (1995) (finding that framing effects similarly influenced lawyer and nonlawyer subjects); Barbara J. McNeil et al., *On the Elicitation of Preferences for Alternative Therapies*, 306 NEW ENG. J. MED. 1259, 1261–62 (1982) (finding that physicians, like nonexperts, are susceptible to framing effects); Gregory B. Northcraft & Margaret A. Neale, *Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions*, 39 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 84, 94–96 (1987) (finding that real estate agents, like nonexperts, are influenced by anchoring effects when estimating real estate prices).

36. Rachlinski et al., *supra* note 7, at 1256 (“With regard to the two phenomena that we have tested in prior studies of generalist judges—anchoring and framing—the bankruptcy judges performed much like the generalist judges we have previously studied.”); *see also* Theodore Eisenberg, *Differing Perceptions of Attorney Fees in Bankruptcy Cases*, 72 WASH. U. L.Q. 979, 983–85 (1994) (reporting evidence suggesting that bankruptcy judges are susceptible to self-serving or egocentric biases when making judgments).

37. Rachlinski et al., *supra* note 7, at 1257.

38. The argument developed here bears some relation to Mark Seidenfeld’s argument that accountability created by the prospect of judicial review might enhance agency decisionmaking. *See* Mark Seidenfeld, *Cognitive Loafing, Social Conformity, and Judicial Review of Agency Rulemaking*, 87 CORNELL L. REV. 486, 509 (2002) (“Judicial review provides accountability, within the core psychological concept . . . .”); *see also* Mark Seidenfeld, *The Psychology of Accountability and Political Review of Agency Rules*, 51 DUKE L.J. 1059, 1061–68 (2001) (reviewing the psychological research on accountability).

39. According to James Flanagan, “the agency’s power to review the findings of the ALJ” is a “fundamental premise of administrative adjudication.” Flanagan, *Update on Developments*,

Professor James Moliterno observes, this review can be quite searching; none of an ALJ's findings, including factual determinations, are binding on the agency because "[t]he agency is omnipotent when it comes to administrative judge decisions."<sup>40</sup> This means that ALJs are accountable to a superior—the agency head or the agency head's designates, like the agency's general counsel's office—in a way that generalist judges are not. And on top of this, ALJs' decisions, like the decisions of generalist judges, are also subject to judicial review in the courts.

The ubiquitous accountability that ALJs face might enhance the quality of their decisionmaking because research shows that accountability often leads to better decisionmaking.<sup>41</sup> When

---

*supra* note 29, at 401; *see also* James E. Moliterno, *The Administrative Judiciary's Independence Myth*, 41 WAKE FOREST L. REV. 1191, 1224 (2006) ("[A]dministrative judges are crucially dependent, both in that their decisions often require executive affirmation and in that they are always subject to executive review . . .").

This "fundamental premise" is embedded in the Administrative Procedure Act, which gave rise to the modern ALJ in federal agencies. *See* 5 U.S.C. § 557(b) (2006); *see also* Flanagan, *Redefining the Role*, *supra* note 29, at 1364. Moreover, the Model State Administrative Procedure Act adopted the federal standard in 1981, as did many of the administrative procedure acts in the states. *Id.*

In recent years, some states—particularly those that have embraced a "central panel" system, *see supra* note 29, have moved to enhance the finality of ALJ decisions, *see, e.g.*, Flanagan, *Redefining the Role*, *supra* note 29, at 1373 (observing that "[a]t the beginning of the 1990s, in almost all the states including those with central panels, the statutes permitted the agency to amend the ALJ's findings of fact and conclusions of law with relative ease," but noting that this began to change in the early 1990s); Jim Rossi, *ALJ Final Orders on Appeal: Balancing Independence with Accountability*, J. NAT'L ASS'N ADMIN. L. JUDGES, Fall 1999, at 1, 2 ("[F]ollowing the proliferation of central panels, many states have increasingly given ALJ orders de jure or de facto finality by taking away an agency's opportunity to review ALJ decisions or requiring agencies to accept ALJ findings unless the agency overcomes a fairly rigorous evidentiary or reasoning burden.").

Despite this innovation in some states, ALJs, even those who preside in central panel states, are much less independent and much more accountable than their counterparts in the judicial branch, as Professor James Moliterno explains:

[T]he central fact of review remains: central panel administrative judges' decisions are also reviewed directly by agencies, often case-party agencies. This is the key fact that largely deprives administrative judges of the judicial independence trait. In reality, nothing is fundamentally changed with the institution of the central panel concept. The administrative judiciary simply becomes a judiciary within the executive branch/administrative state. This feature may further serve to insulate the administrative judge in an impartiality sense, but the administrative judge remains a member of the executive branch, fully subject to override by the agency.

Moliterno, *supra*, at 1233.

40. Moliterno, *supra* note 39, at 1225.

41. *See, e.g.*, Jennifer S. Lerner & Philip E. Tetlock, *Accounting for the Effects of Accountability*, 125 PSYCHOL. BULL. 255, 260–62 tbl.1 (1999) (documenting the extensive literature on accountability).

decisionmakers know the views of the parties to whom they are accountable, they generally anticipate and adopt those views.<sup>42</sup> This suggests that when ALJs know how their agency head would decide the case before them, they are likely to decide in the manner that they believe their agency head would decide, hewing closely to the agency's interpretation of its governing law. When decisionmakers do not know the views of the parties to whom they are accountable, they are more likely to engage in "self-critical" and "effortful" processing, often leading to improved decisionmaking.<sup>43</sup> This suggests that when ALJs know their decisions will be reviewed, but are uncertain how their agency head would decide the case, they will exert more effort and care than is customary, perhaps leading to more rational and deliberative decisionmaking. In either case, then, there is reason to believe that ALJs, at least relative to their counterparts in the judiciary, might make better decisions because they have a higher degree of accountability.<sup>44</sup>

As with expertise, the apparent difference in accountability between ALJs and generalist judges is easy to overstate. For example, both generalist judges and ALJs are similarly accountable in the sense that both offices tend to carry a high degree of job security. ALJs are certainly less secure than Federal District Judges, but they are apt to have more job security than state judges and are likely to have some

---

42. See, e.g., Lerner & Tetlock, *supra* note 41, at 256 ("When audience views are known prior to forming one's own opinion, conformity becomes the likely coping strategy. People can simply adopt positions likely to gain the favor of those to whom they are accountable, thereby allowing them to avoid the unnecessary cognitive work of analyzing the pros and cons of alternative courses of action, interpreting complex patterns of information, and making difficult trade-offs." (citations omitted)); Philip E. Tetlock, *The Impact of Accountability on Judgment and Choice: Toward a Social Contingency Model*, 25 *ADV. EXPERIMENTAL SOC. PSYCHOL.* 331, 340-41 (1992) (observing that when people know the views of those to whom they are accountable, they frequently adopt those views).

43. Lerner & Tetlock, *supra* note 41, at 259, 260-62 tbl.1 (describing this phenomenon and identifying the decision heuristics and biases that are attenuated, as well as a small number that are actually amplified, by this form of accountability); see also Tetlock, *supra* note 42, at 343 ("Predecisional accountability to unknown audiences frequently motivates vigilant, complex, and self-critical thinking."). For refinements, see Itamar Simonson & Peter Nye, *The Effect of Accountability on Susceptibility to Decision Errors*, 51 *ORGANIZATIONAL BEHAV. & HUMAN DECISION PROCESSES* 416, 437 (1992). "[A]ccountability is expected to reduce errors if decision makers can identify the response that would be regarded as more rational, but that response is different from the one that unaccountable decision makers tend to make." *Id.*

44. For the argument that any behavioral analysis of legal decisionmaking should consider the potential impact of accountability on the decisionmaker, see Gregory Mitchell, *Why Law and Economics' Perfect Rationality Should Not Be Traded for Behavioral Law and Economics' Equal Incompetence*, 91 *GEO. L.J.* 67, 110-14 (2002).

measure of civil service protection or the like. Job security is obviously one important aspect of accountability, but it is not what we are focusing on here. The psychological evidence on accountability suggests that scrutiny of individual decisions by superiors, as opposed to lengthy performance reviews, can reduce reliance on intuitive judgment. It is in this sense that the ALJs are somewhat more accountable than generalist trial judges. Many of their decisions are merely recommendations to agencies, which may or may not be adopted.<sup>45</sup> To be sure, in some cases, these recommendations are adopted with sufficient frequency that they might as well be final orders, but even if an ALJ's recommendations as to a final order are accepted, they are still subject to review by a trial or appellate court, meaning that ALJs are subject to at least one additional layer of review that trial judges are not. Therefore, ALJs, particularly those who issue only recommendations, make decisions that are more apt to be scrutinized in the way that psychologists contend might reduce reliance on intuitive processing.<sup>46</sup>

### C. Feedback

The third reason we expected to find that ALJs might make more rational and deliberative decisions is a consequence of the second. Because ALJs are subject to review by the agencies that employ them (as well as by the courts in those cases subject to judicial review), they receive much more *feedback* on the accuracy of their

---

45. For example, ALJs in Oregon might face the same degree of scrutiny as the trial court: Ninety eight percent of the OAH's orders are final orders. These orders generally cannot be changed by agencies after ALJs issue them. They are appealable to a circuit court or the Court of Appeals (unemployment insurance decisions are appealable to the Employment Appeals Board). The remaining OAH orders are proposed orders. Proposed orders are decisions recommended to agencies by ALJs based on their review of the facts and law. Agencies are not required to accept the recommendations. However, if they do not, they are required to explain the reason for all "substantial" changes. If findings of fact are changed, they must show how most of the evidence at hearing supports their version of the facts; on appeal to the Court of Appeals, the Court can look at the entire record and determine independently (*de novo*) whether it agrees with the agency's version.

Office of Admin. Hearings, The OAH, [http://www.oregon.gov/OAH/The\\_OAH2.shtml](http://www.oregon.gov/OAH/The_OAH2.shtml) (last visited Mar. 1, 2009).

46. The fact that some ALJs make only recommended decisions might encourage reliance on intuitive processes. Knowing that a decision might be reviewed by an agency head could lead to a diffusion of responsibility so that ALJs worry less about the accuracy of their decisions. Although this is possible, we doubt it. We suspect that ALJs, like most judges, take their positions very seriously and use whatever resources and information that they have available to try to make sound judgments.

decisions than does a typical judicial-branch judge. In contrast to generalist judges, who face appellate review in only a very small number of cases<sup>47</sup> and on limited bases,<sup>48</sup> ALJs receive review much more commonly. Some ALJs only make recommendations to agencies, which might mean that they receive feedback in virtually every case. In the large majority of cases in which agencies embrace the ALJs' decisions,<sup>49</sup> the ALJs learn that their decisions were accurate, at least from the perspective of the agency employing them.<sup>50</sup> And in the cases in which the agency rejects their decisions, the ALJs receive immediate feedback regarding the flaws in their decisionmaking, at least from the perspective of the agency. Furthermore, some agencies conduct routine audits of ALJ decisions, producing a kind of report card for each judge.<sup>51</sup>

Psychologists have found that feedback—if it is provided promptly and offers meaningful insight into the causes and consequences of decisionmaking flaws—can improve decisionmaking.<sup>52</sup> This is precisely the kind of feedback that ALJs, in

---

47. See, e.g., C.K. ROWLAND & ROBERT A. CARP, *POLITICS AND JUDGMENT IN FEDERAL DISTRICT COURTS* 8 (1996) (“[O]nly about 20 percent of all district court cases are appealed in any given year.”); Theodore Eisenberg, *Appeal Rates and Outcomes in Tried and Nontried Cases: Further Exploration of Anti-Plaintiff Appellate Outcomes*, 1 J. EMPIRICAL LEGAL STUD. 659, 685 (2004) (“About 20 percent of cases with definitive trial court judgments generate appeals, with tried cases appealed at about twice the rate of nontried cases.”).

48. See, e.g., ROWLAND & CARP, *supra* note 47, at 9 (observing that interpretations of fact are free from review); William M. Landes & Richard A. Posner, *Harmless Error*, 30 J. LEGAL STUD. 161, 162 (2001); Maurice Rosenberg, *Standards of Review*, in *RESTRUCTURING JUSTICE: THE INNOVATIONS OF THE NINTH CIRCUIT AND THE FUTURE OF THE FEDERAL COURTS* 30, 31 (Arthur D. Hellman ed., 1990) (explaining bases of review).

49. See, e.g., Charles E. Daye, *Powers of Administrative Law Judges, Agencies, and Courts: An Analytical and Empirical Assessment*, 79 N.C. L. REV. 1571, 1616 (2001) (reporting, based on an empirical examination of ALJ decisionmaking in North Carolina, that agencies fully adopted 82 percent and partially adopted 6 percent of the decisions recommended by their ALJs).

50. We do not address the issue of political conflicts between the ALJs and the agency. An ALJ could interpret the law and facts perfectly well, but face an agency bent on moving the law in a new direction, or implementing a new policy. Although such conflicts would produce reversals of the ALJ, they are not really errors.

51. The Social Security Administration, for example, conducts these kinds of audits. U.S. GEN. ACCOUNTING OFFICE, *SSA DISABILITY DECISION MAKING: ADDITIONAL STEPS NEEDED TO ENSURE ACCURACY AND FAIRNESS OF DECISIONS AT THE HEARINGS LEVEL 9* (2003), available at <http://www.gao.gov/new.items/d0414.pdf> (describing the SSA's Office of Quality Assurance and Performance Assessment's process for conducting quality assurance reviews of a random sample of ALJ decisions). This is similar to the peer review we recommended in a previous paper. Guthrie et al., *Blinking*, *supra* note 8, at 39.

52. See, e.g., ROBIN M. HOGARTH, *EDUCATING INTUITION* 88–90 (2001) (distinguishing “kind” environments, which provide valuable feedback and in turn can improve intuitions, from “wicked” environments, which do not).

contrast to generalist trial judges, typically receive. The availability of this kind of feedback thus provides yet another reason to be optimistic about the quality of ALJ decisionmaking.

To be sure, if the decisions ALJs make are adopted routinely, they will not receive much better feedback than trial judges.<sup>53</sup> Also, the feedback that agencies give to ALJs might be unclear or muted. But the average ALJ will tend to get more feedback from a reviewing agency (and subsequent judicial review) than the average trial judge will. Whether such feedback is sufficient to induce ALJs to rely less on intuition is unclear.

#### D. Summary

ALJs enjoy some potential advantages over generalist trial judges that might enable them to make more rational decisions. They often specialize in limited areas of law, their decisions are more apt to be scrutinized on review, and they receive more frequent and prompt feedback. For these three reasons—specialization, accountability, and feedback—we hypothesized that ALJs might make predominantly rational and deliberative, rather than predominantly intuitive and impulsive, decisions.<sup>54</sup>

### III. STUDYING THE ADMINISTRATIVE LAW JUDGE

We have reported evidence in our prior work that supports our intuitive-override model of judging.<sup>55</sup> That work, however, is based on

---

53. Adoption rates vary. In Oregon, adoption rates exceed 98 percent in many categories of cases. See Office of Admin. Hearings, *supra* note 45. Arizona reports that adoption rates of ALJ decisions are over 85 percent. OFFICE OF ADMIN. HEARINGS, THIRTEENTH ANNUAL REPORT 1 (2008), available at <http://www.azoah.com/13thAnnualReport.pdf> (reporting that the Arizona agency acceptance of central panel ALJs' findings of fact and conclusions of law without modification was 92.6 percent, and that the overall rate of adoption of entire ALJ decisions (that is, including the remedy portion of the recommendation) without modification was 86.64 percent).

54. We recognize, of course, that some previous empirical studies of ALJs suggest that they are not ideal decisionmakers. Some ALJs have been shown to produce decisions that vary by the race of the litigants, see Stephen Labaton, *Benefits Are Refused More Often to Disabled Blacks, Study Finds*, N.Y. TIMES, May 11, 1992, at A1 (summarizing a 1992 study by the General Accounting Office and reporting that “[f]or 30 years, blacks with serious ailments have been much more likely than whites to be rejected for benefits under Social Security disability programs”), and others have been shown to be wildly erratic, see Jaya Ramji-Nogales, Andrew I. Schoenholtz & Philip G. Schrag, *Refugee Roulette: Disparities in Asylum Adjudication*, 60 STAN. L. REV. 295, 296 (2007) (reporting “amazing disparities” in asylum grant rates among immigration judges).

55. See sources cited *supra* note 8.



evidence gathered from trial judges who are generalists (and, in one instance, from a sample of federal bankruptcy judges). The purpose of this Article is to explore whether this model also captures the way ALJs appear to make decisions. Because they are similar to generalist judges in education, experience, and role, it seems likely that ALJs would behave in much the same way as generalist judges. On the other hand, as explained above, ALJs might have some decisionmaking advantages, given their subject-matter expertise, accountability to agency superiors, and more frequent receipt of timely feedback.

To explore ALJ decisionmaking, we conducted two lines of investigation. First, we explored whether ALJs appear more inclined toward intuition or deliberation when responding to a set of generic problems that make up the “Cognitive Reflection Test” (CRT).<sup>56</sup> Second, we asked ALJs to respond to a series of hypothetical, judicial-decisionmaking problems in which we explored whether different variables triggered intuitive decisionmaking. In this Part, we explain our methods and then present our results.

#### A. *Methods*

To explore ALJ decisionmaking, we gave questionnaires to two groups of ALJs attending judicial education conferences at which we appeared as speakers. The first group, which consisted of thirty-nine ALJs, attended the City of New York Administrative Judicial Institute in New York City in June of 2008 (the “city conference”). The second group, which consisted of 103 ALJs, attended the National Association of Administrative Law Judiciary 2008 Annual Conference in New York City in October of 2008 (the “national conference”).

Among the ALJs who attended the city conference, twenty-two were men and seventeen were women. They had an average of 10.1 years of experience, with a range of zero to twenty-nine years. In response to a question concerning their affiliation with a political party, thirty-one identified themselves as Democrats, three as Republicans, and five declined to answer this question. Among the ALJs who attended the national conference, forty were men, fifty-seven were women, and six declined to disclose their gender. These judges had an average of 10.8 years of experience, with a range of

---

56. *See infra* Part II.B.

zero to thirty-one years' experience. They also identified mostly as Democrats (seventy-six), as opposed to Republicans (eleven), although sixteen did not respond to this question.

The judges in the study had a vast and varied range of judicial experience and expertise. We asked the judges to identify "what type of cases you normally hear." The ALJs who attended the city conference gave the following responses: employee discipline, including police matters (nine judges); administrative code violations (six judges); environmental control (five judges); health, safety, building code (five judges), special education (five judges); traffic and parking (four judges); vehicle forfeiture (four judges); discrimination claims (two judges); Section 8 (housing) claims (two judges); taxi and limousine commission (two judges); and one judge each in contract cases, licensing, sanitation, tax, department of transportation (which might have been taxi and limo), and zoning.<sup>57</sup>

The ALJs who attended the national conference included an even more diverse array of judges. These judges traveled to the conference from fourteen states: New York (forty-seven judges); District of Columbia (nine judges); California (four judges); Georgia (seven judges); Illinois (two judges); Iowa (two judges); Kentucky (one judge); Maryland (five judges); North Carolina (two judges); South Carolina (one judge); Texas (two judges); Vermont (one judge); Washington (two judges); West Virginia (two judges). In addition, we had two ALJs from Quebec and one federal magistrate judge. These ALJs reported presiding over a wide range of cases as well: employment, including unemployment insurance (seven judges); professional licensing (seven judges); worker's compensation (seven judges); special education (six judges); housing (section 8 and rent control) (six judges); environmental (six judges); traffic (including parking and DUI) (five judges); antidiscrimination (or human rights) (four judges); health code (four judges); medical/healthcare (including Medicare) (four judges); consumer protection (four judges); public works (three judges); building permits (two judges); labor standards (two judges); child abuse/foster care (two judges); campaign finance (two judges); parole violations (two judges); disability (two judges); and one judge each in insurance, immigration, tax, liquor, securities, forfeiture, and taxi and limousine. Additionally, twenty-one judges indicated that they were part of the state's "central

---

57. Note that some of the judges identified more than one type of case in response to this question.

panel” and hence were available for all kinds of cases. Another eleven did not respond to this question.

We also asked this group of judges to identify how their decisions are reviewed. Specifically, we asked whether they make “recommendations that are either adopted, rejected, or modified by an agency” or whether their decisions are simply subject to review by an agency or a court. Of the ninety-four judges who answered the question, 44.7 percent (forty-two judges) indicated that they make recommendations, 39.3 percent (thirty-seven judges) indicated that they make decisions that are subject to review by an agency, and 51.1 percent (forty-eight judges) reported making decisions that are subject to review by a court. As the numbers indicate, some of the judges checked more than one box: two checked both recommendation and review by agency; three checked recommendation and review by court; eight checked both review by agency and review by court; and ten checked all three. To facilitate some of our analyses, we divide the judges into the 44.7 percent who indicated that they make recommendations in at least some instances and the 55.3 percent who are only subject to review by either an agency or a court.<sup>58</sup>

We followed the same procedure at both conferences. Before beginning our formal presentation, we randomly distributed questionnaires to the ALJs in person. We asked the ALJs to read and respond to each of the questions and to provide some basic demographic information. All of the questionnaires included the CRT as well as a variety of problems designed to test judicial decisionmaking.<sup>59</sup> We informed the ALJs that participation was entirely voluntary, that they could continue or cease participating at any time, and that their responses would be anonymous. We also gave the ALJs at each conference the opportunity to limit our use of their results to the conference itself. None of the ALJs at the city conference did this, but four of the judges at the national conference exercised this option, and their results are excluded from our analysis.

---

58. We do not have this data available for the city conference judges (although most are in the third category), so we do not categorize them in this way or include them in specialized analyses that distinguish the ALJs from one another based on the type of review to which their decisions are (or are not) subjected.

59. Copies of the original versions of these problems are included in Appendices A and B, along with a copy of the questions that elicited demographic information about the judges.

## B. Results: Cognitive Reflection

We first explored whether ALJs, like generalist trial judges, solve generic problems in an intuitive or deliberative way. To do so, we gave the ALJs the CRT.<sup>60</sup>

1. *Cognitive Reflection Test.* Created by Frederick, the CRT is a simple, three-item test designed to distinguish intuitive and impressionistic processing from deliberative and deductive processing.<sup>61</sup> More specifically, the CRT measures “cognitive reflection,” which is “the ability or disposition to resist reporting the response that first comes to mind.”<sup>62</sup>

The CRT is printed below, as Figure 1. Each of the three CRT items has an intuitive answer that almost immediately comes to mind, but the intuitive answer is wrong. The correct answer is fairly easy to determine, but ascertaining it requires deliberation. As Frederick explains, the CRT items are simple because “their solution is easily underst[andable] when explained, yet reaching the correct answer often requires the suppression of an erroneous answer that springs ‘impulsively’ to mind.”<sup>63</sup>

Figure 1. *Cognitive Reflection Test*<sup>64</sup>

<p>(1) A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost? _____cents</p> <p>(2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____minutes</p> <p>(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____day</p>
---

By way of illustration, consider the “bat-and-ball” problem, which is the first problem in Figure 1. When asked to solve this problem, most people immediately report “ten cents” as their

60. Frederick, *supra* note 16, at 26–28.

61. *Id.* at 27.

62. *Id.* at 35.

63. *Id.* at 27.

64. Researchers utilized the same three questions in Frederick’s study. *Id.*

answer.<sup>65</sup> Although this response is intuitive, and although it seems as though it must be correct, a little deliberation reveals that it is inaccurate. If the ball costs ten cents, as most people think when they first encounter the problem, and the bat costs one dollar more, as the problem specifies, the bat must cost \$1.10. This means that the total cost of the bat and ball is \$1.20 rather than \$1.10 as specified by the problem. The correct answer is five cents—that is, the ball costs five cents, the bat costs \$1.00 more or \$1.05, and the two together cost \$1.10, as required by the problem. It is an easy problem, but to get the right answer, respondents have to ignore the initial, intuitive, incorrect answer that occurs to them and think through the problem in a deliberative way.

This is true for the second CRT problem—the “widget” problem—as well. On this problem, the answer that comes immediately to mind for most people is one hundred minutes.<sup>66</sup> The correct answer, however, is five minutes. If, as the problem specifies, five machines make five widgets in five minutes, it takes each machine five minutes to make a widget. Thus, one hundred machines would make one hundred widgets in five minutes (and five hundred machines would make five hundred widgets in five minutes, one thousand machines would make one thousand widgets in five minutes, etc.).<sup>67</sup>

The third problem—the “lily-pad” problem—also invites an intuitive and inaccurate response. For most people, the answer that immediately comes to mind is twenty-four days.<sup>68</sup> The correct answer on this problem, as on the others, is obvious upon reflection. If, as the problem specifies, the patch of lily pads doubles each day and covers the entire lake on the forty-eighth day, it must cover half of the lake a day earlier. This means that the correct answer is forty-seven days, not twenty-four days.

For those who suppress their impulsive responses and deliberate, the CRT items are simple and straightforward. Most people are unable, or perhaps unwilling, to overcome their impulsive reactions to the problems. In thirty-five studies involving thousands of

---

65. *Id.* at 26–27.

66. *Id.* at 27.

67. As we have noted elsewhere, see Guthrie et al., *Blinking*, *supra* note 8, at 11 n.57, this problem assumes that each machine produces widgets at the same rate, which seems a reasonable assumption given the structure of the problem and of the CRT as a whole.

68. Frederick, *supra* note 16, at 27.

respondents, Frederick found that subjects correctly answered, on average, only 1.24 of the three items.<sup>69</sup> Among all of the participants in Frederick's studies, only 17 percent answered all three questions correctly; roughly twice that many (33 percent) got all three problems wrong.<sup>70</sup>

The CRT illustrates the predominance of intuitive over deliberative processing in three ways. First, most people perform poorly on the CRT, even though the problems are simple if approached deliberately.<sup>71</sup> Ironically, people tend to perform better on similar but *harder* problems that signal that deliberation is required.<sup>72</sup> Frederick explains, for example, that people "miss the 'bat and ball' problem far more often than they miss the 'banana and bagel' problem: 'A banana and a bagel cost 37 cents. The banana costs 13 cents more than the bagel. How much does the bagel cost?'"<sup>73</sup>

Second, among the infinite number of potential inaccurate responses to each of the CRT problems, the intuitive answers identified above—ten cents in the bat-and-ball problem, one hundred minutes in the widget problem, and twenty-four days on the lily-pad problem—are the most commonly offered responses.<sup>74</sup>

Third, and perhaps most impressively, subjects who select the inaccurate, intuitive response are more likely than subjects who answer the problems correctly to indicate that the problems are easy.<sup>75</sup> In his administration of the bat-and-ball problem, for example, Frederick found that subjects who provided the intuitive, but inaccurate, response—that is, ten cents—predicted that 92 percent of people would solve the problem correctly because it seemed quite easy to them. Among the subjects who actually provided the correct answer, however, only 62 percent predicted that other people would solve the problem correctly.<sup>76</sup> Having suppressed their intuitive

---

69. *Id.* at 28, 29 tbl.1.

70. *Id.*

71. *See id.* at 27.

72. *Id.* at 28.

73. *Id.* In contrast to the bat-and-ball problem, the banana-and-bagel problem does not suggest an intuitive response. Accordingly, most people attempt to solve the problem deliberately by explicitly or implicitly using algebra—i.e.,  $x = \text{bagel}$ ,  $y = \text{banana}$ ,  $y = x + 13$ , so  $x + x + 13 = 37$ . Solving for  $x$ ,  $x = 12$  and  $y = 25$ .

74. *Id.* at 27.

75. *Id.*

76. *Id.*

reaction in favor of deliberation, they appreciated that the problem, though simple if one deliberated, was actually tricky.

Although the CRT correlates with standard assessments of intelligence,<sup>77</sup> it does measure what an IQ test measures. Rather, it tests a respondent's ability to suppress intuition in favor of deliberation in a setting where intuition is misleading.<sup>78</sup> This is why we find it to be of interest for judges, who must commonly suppress their intuition in favor of following a deliberative decisionmaking process. Furthermore, people who score well on the CRT also appear to resist falling prey to a range of common cognitive errors in judgment, including the conjunction fallacy, imperfect Bayesian reasoning, overconfidence, hyperbolic discounting, and risk preferences resulting from framing effects.<sup>79</sup> Many of these errors can influence judgments in legal contexts, and we have found that generalist judges are vulnerable to some of them. The CRT might be able to identify judges who suppress the intuitive processes that can produce these errors and rely instead on deliberative processes.

2. *CRT and ALJs.* Curious about how generalist judges would respond to the cognitive reflection test, we gave it to nearly half of the sitting trial judges in Florida at a judicial education conference. The judges obtained an average score of 1.23,<sup>80</sup> slightly higher than the average score of students at the University of Michigan (1.18) and slightly lower than the average score of students at Harvard (1.43).<sup>81</sup> Like other subjects, the generalist judges tended to select the intuitive, but incorrect, answer on the problems,<sup>82</sup> and like other subjects, the judges who did so tended to think that the problems were easier than did those judges who answered the problems correctly.<sup>83</sup>

The results seem predictable—judges performed about as well as the most educated adults on the CRT. The results were not foreordained, however. Judges spend an enormous amount of time

---

77. *Id.* at 35.

78. *Id.*

79. Jorg Oechssler, Andreas Roeder & Patrick W. Schmitz, *Cognitive Abilities and Behavioral Biases* 4–6 (Inst. for the Study of Labor, Discussion Paper No. 3481, 2008), available at <http://ssrn.com/abstract=1294555>.

80. Guthrie et al., *Blinking*, *supra* note 8, at 14.

81. Frederick, *supra* note 16, at 29 tbl.1.

82. Guthrie et al., *Blinking*, *supra* note 8, at 16.

83. *Id.*

sorting through and applying rules that might conflict with their intuitive responses. As Table 1 indicates, undergraduates at MIT, who largely specialize in math, engineering, and science, perform much better on the CRT. The sense that intuition should be ignored, or at least checked by logic or computation, is well developed in such professions, and it shows on the CRT results. And that is all to the good. No one wants to drive over a bridge that was designed by a civil engineer who relied on intuition as to whether it had enough structural support; we want someone who did the math. Similarly, we might want judges who have habits of mind that lead them to check their intuitive responses with deliberation, but this is not what we found among generalist judges on the CRT.

ALJs might be different, as we have discussed. To see how the ALJs would perform on the CRT, we gave the CRT to both samples of ALJs. We excluded the three judges from the city conference and the thirteen judges from the national conference who did not complete all three questions.<sup>84</sup> Among the remaining judges, the city conference attendees obtained an average score of 1.50, and the ALJs at the national conference obtained an average score of 1.27.<sup>85</sup> Collectively, the ALJs obtained an average score of 1.33, which is roughly comparable to the score obtained by the Florida state trial judges.<sup>86</sup> Overall, 30.2 percent provided incorrect answers on all three items; 27.8 percent answered one item correctly; 20.6 percent answered two items correctly; and 21.4 percent got all answered all three items correctly. Among those who provided inaccurate responses, the most commonly offered responses on the CRT problems were the intuitive but inaccurate responses suggested by the problem. On the bat-and-ball problem, for example, 93.7 percent of

---

84. Among these sixteen judges, five answered none of the questions; four answered both of the first two questions with the intuitive (but wrong) answer, and did not answer the third question; one answered the first question with the correct answer, the second question with the intuitive (but wrong) answer, and did not answer the third question; one answered the first question with the intuitive (but wrong) answer and did not respond to the second or third questions; one answered the first question with the intuitive answer and the second question with a wrong (but not the intuitive) answer, and did not respond to the third question; one answered the first question with the intuitive but incorrect answer, the second question with the correct answer, and did not respond to the third question; three answered the first question with the intuitive (but wrong) answer, skipped the second question, and answered the third question with the correct answer.

85. An ordered logit regression revealed that the two groups did not differ significantly.  $z = 1.08, p = .28$ .

86. An ordered logit regression revealed that the ALJs did not perform significantly differently than the Florida trial judges.  $z = .80, p = .45$ .



those who responded inaccurately gave the intuitive response suggested by that problem; on the widget problem, 52.7 percent of those who got the question wrong gave the intuitive but inaccurate response; and on the lily-pad problem, 64.9 percent of those who got the question wrong gave the intuitive but inaccurate response. As depicted in Table 1, the ALJs performed much like trial judges and many college students.

*Table 1. Overall CRT Results: Judges and Selected Others*<sup>87</sup>

Subject Population (and Sample Size)	Mean	Percent 0 Right	Percent 1 Right	Percent 2 Right	Percent 3 Right
MIT (61)	2.18	7	16	30	48
Carnegie Mellon (746)	1.51	25	25	25	25
Harvard (51)	1.43	20	37	24	20
ALJs (126)	1.33	30	28	21	21
Florida Judges (252)	1.23	31	31	24	15
Michigan/Ann Arbor (525)	1.18	31	33	23	14
Bowling Green (52)	.87	50	25	21	6
Michigan State (118)	.79	49	29	16	6
Toledo (138)	.57	64	21	10	5

The ALJs, like the other respondents, provided predominantly intuitive responses to the CRT items, though most of them demonstrated that they could override their intuition in at least some circumstances. Nonetheless, the CRT results show that ALJs, like generalist judges, gravitate toward intuitive rather than deliberative processing.<sup>88</sup>

87. All of the nonjudge data on this table were collected on college students, and comes from Frederick, *supra* note 16, at 29 tbl.1. The data from the Florida Judges come from Guthrie et al., *Blinking*, *supra* note 8, at 15 tbl.2. Note that the table in our previous article mistakenly reports the sample size as 192 rather than 252.

88. Previous research with other populations shows that men perform slightly better than women on the CRT. Frederick, *supra* note 16, at 37–38, 38 tbl.6 (showing male subjects with a mean score of 1.47 and female subjects with a mean score of 1.08). We did not find this to be the case with the Florida trial judges in our previous work, but we did see differences among the ALJs. The fifty-eight male judges who both completed the CRT and identified their gender scored an average of 1.52, whereas the sixty-six women judges scored an average of 1.17. This difference was marginally significant, per an ordered logit regression, revealing  $z = 1.80$ ,  $p = .07$ .

The average CRT scores among the thirty-seven judges in the national conference sample who make recommendations and completed the CRT was 1.30, as compared to 1.20 among the forty-nine judges who are only subject to review; this was not a significant difference. The analysis was done with an ordered logit regression, revealing  $z = .37$ ,  $p = .72$ . This last result is perhaps no surprise, but is inconsistent with the hypothesis that judges who make

### C. Results: Judicial Decisionmaking

The CRT results suggest that ALJs, like their colleagues on generalist courts, tend to make *ordinary judgments* in a predominantly intuitive way. Although suggestive, these results do not necessarily reveal how ALJs make *judgments on the bench*. To explore whether ALJs make intuitive judgments not only when faced with mathematically oriented word problems, but also when confronted with the kinds of judgment tasks they face on the job, we gave the ALJs who participated in our studies several judicial-decisionmaking problems. More specifically, we tested for susceptibility to six well-known psychological phenomena that are likely to trigger intuitive thinking: anchoring, framing, conjunction fallacy, outcome bias, disregarding, and egocentric bias. We found that ALJs, like generalist judges, are often, though not always, influenced by these phenomena. Our results support our conclusion that ALJs, like generalist trial judges, approach legal problems in a predominantly intuitive, though occasionally deliberative, way.

1. *Anchoring*. When people make numerical estimates, they tend to rely heavily on the first number available to them. The initial number, in other words, provides an “anchor” that exerts disproportionate influence on the estimation process.<sup>89</sup> Psychologists have found, for example, that subjects provided higher estimates of the average temperature in San Francisco if first asked to indicate whether it was higher or lower than 558 degrees, and provided higher estimates of the average price of a college textbook if first asked whether the average price was higher or lower than \$7,128.53.<sup>90</sup> Psychologists refer to this phenomenon as “anchoring.”<sup>91</sup>

Anchoring may be ameliorated by accountability. In research on anchoring, psychologist Phil Tetlock has found that anchoring has less influence on people when they learn they must explain their estimates and that their explanations will be held up for public scrutiny.<sup>92</sup> This

---

recommendations develop radically different habits of mind that lead them to engage in deliberative thinking across the board.

89. See Tversky & Kahneman, *supra* note 9, at 1128 (“[D]ifferent starting points yield different estimates, which are biased toward the initial values.”).

90. PLOUS, *supra* note 35, at 146 (reporting the results of an unpublished study conducted by George Quattrone and his colleagues).

91. Tversky & Kahneman, *supra* note 9, at 1128 (introducing the phenomenon).

92. See Lerner & Tetlock, *supra* note 41, at 262–63 (“Moreover, two especially pervasive tendencies, (a) anchoring on an initial value and insufficiently adjusting a numerical estimate up

manipulation mimics the conditions under which judges sometimes operate, yet we have found that anchors influence the judgments of generalist trial judges in legal settings.<sup>93</sup> Specifically, we have found that anchors influence estimates of damage awards by trial judges.<sup>94</sup> We have found this effect even when the anchors were clearly irrelevant and clearly inadmissible.<sup>95</sup> Because ALJs are accustomed to review, and because their decisions are commonly afforded less deference than those of trial judges, they might be less susceptible to anchoring than their counterparts in the judicial branch.

To explore whether anchoring influences ALJs, we gave the ALJs who attended the national conference a problem we called “*Veronica v. AAC*.” We asked the ALJs participating in our study to imagine that they were presiding over an employment discrimination case brought by the City Commission on Human Rights on behalf of Veronica Sanchez against a company called Administrative Assistants on Call (AAC). In the suit, the ALJs learn, the Commission alleges that AAC violated the administrative code by terminating Veronica, a Mexican-American immigrant, on the basis of her “actual or perceived” race, color, national origin, and citizenship status.

Veronica had been a top employee at AAC, a company that provides off-site secretarial services to small businesses, until a new manager arrived. The new manager allegedly assigned her solely to undesirable projects; used racially offensive terms in her presence; told her to “go back to Mexico”; and accused her of stealing work from “real Americans.” On one occasion, he made these comments in front of Veronica’s daughter, who was visiting her at work, and in front of one of Veronica’s coworkers, who corroborated Veronica’s testimony. Veronica complained, and the manager proceeded to fire her.

AAC does not deny Veronica’s allegations, but it challenges the Commission’s request for compensatory damages on Veronica’s behalf because Veronica found comparable work immediately after her termination and because AAC terminated the manager’s employment.

---

or down from that anchor and (b) weighting sunk costs when considering future investments are also reduced by accountability.” (citations omitted)).

93. See Guthrie et al., *Judicial Mind*, *supra* note 8, at 787–92; Wistrich et al., *supra* note 8, at 1286–91.

94. Guthrie et al., *supra* note 8, at 791–92.

95. Wistrich et al., *supra* note 8, at 1288–91.

The ALJs learn that the administrative code authorizes them to award compensatory damages for “mental anguish.” They also learn that controlling case law provides that such an award is justified when “there is credible testimony of some degree of emotional distress.” In this instance, we informed the ALJs that Veronica provided credible testimony that she has suffered from “anxiety, sleeplessness, and bad dreams.” We also informed the ALJs that Veronica mentioned, as an aside, that she recently saw a court TV show featuring a case she claimed was similar to hers.

Unbeknownst to the ALJs, we assigned roughly half of them to a control group and half to an anchor group. The ALJs in the control group learn that the claimant in the court TV show Veronica watched received a compensatory damage award, but no dollar amount was specified. The ALJs assigned to the anchor group learn that Veronica mentioned that the claimant in the court TV show had received a compensatory damage award in the amount of \$415,300. We then asked the ALJs in both groups to indicate the amount of compensatory damages they would award Veronica for her mental anguish. The anchor in this instance—an alleged award observed on a court TV show—was irrelevant to the ALJs’ determination under the law.

Furthermore, we directly tested whether accountability might influence the ALJs’ judgments by attempting to replicate Tetlock’s finding that making people explain their decisions would reduce or eliminate the anchoring effect. We informed roughly half of the judges that further review of their award was “extremely unlikely.” Judges in these conditions were simply told to assign an award. We informed the other judges that review of their award and their explanation was very likely. These judges were asked both to assign an award and to provide an explanation for the award.

The judges thus evaluated one of four different versions of the problem: (1) no anchor and no explanation required; (2) no anchor and an explanation required; (3) anchor and no explanation required; and (4) anchor and an explanation required.<sup>96</sup>

---

96. Although we were concerned that the request for an explanation might have induced some judges not to answer the question, this does not appear to have been the case. Twenty-one judges did not respond to this question: seven in the control, no-explanation condition; four in the control, explanation condition; four in the anchor, no-explanation condition; and six in the anchor and explanation condition. Among the forty-three judges who provided an award in the explanation conditions, four did not provide an explanation (one in the control condition and three in the anchor condition). These judges were nevertheless kept in the analysis.

The anchor affected the ALJs' judgment. The judges in the control awarded an average amount of \$35,488; the judges in the anchor group awarded a much larger \$58,775 on average. The averages are a bit misleading because the distribution of damage awards in both conditions was highly positively skewed, but as Table 2a shows, the anchor shifted the entire distribution of awards upward. A sizeable number of judges thought that no damage award was appropriate, even though we had indicated that the employer was only contesting the "request for damages." In the control condition, 28.6 percent (twelve out of forty-two) of the judges provided an award of \$0, whereas only 7.5 percent (three out of forty) did so in the anchor condition.

*Table 2a. Damage Awards by Condition in the Anchoring Problem*

Condition (and sample size)	Average Award (\$)	Percent \$0 Award	25th Percentile	Median	75th Percentile
No Anchor (42)	35,488	28.6	0	6,250	10,000
Anchor (40)	58,775	7.5	5,000	50,000	100,000

Accountability did not affect judgment by itself. The judges who were not asked to provide an explanation gave an average award of \$66,308, whereas the judges who were asked to provide an explanation gave an average damage award of \$29,198. On the face of it, it seems that the explanation requirement reduced awards substantially. But as Table 2b shows, the explanation did not have much effect on the overall range of awards, suggesting that the differences in the averages are the result of a few extremely high awards in the no-explanation condition. In fact, the no-explanation condition included a \$400,000 award and a \$1,000,000 award, whereas the highest award in the explanation condition was a \$200,000 award.<sup>97</sup> Also, in the no-explanation condition, 12.8 percent (five out of thirty-nine) of the judges provided an award of \$0; in the explanation condition, by contrast, 23.3 percent (ten out of forty-three) did so.

---

97. The \$1,000,000 award was in the no anchor condition. The five next highest awards were in the anchor condition.

Table 2b. *Damage Awards by Explanation in the Anchoring Problem*

Condition (and sample size)	Average Award (\$)	Percent \$0 Award	25th Percentile	Median	75th Percentile
No Explanation (39)	66,308	12.8	5,000	10,000	50,000
Explanation (43)	29,198	23.3	2,000	10,000	50,000

Even though the explanation had no effect on its own, we were primarily interested in whether requiring an explanation would ameliorate the effect of the anchor. Thus, we assessed the interaction between the anchor and the explanation. Table 2c reports these results. As the sample sizes get small, the results become somewhat erratic. But a comparison of the anchor and control conditions, both with and without explanations, suggests that the anchoring effect is robust; requiring an explanation did not diminish its power much, if at all. The magnitude of the anchoring effect, as measured by the difference between the averages in the anchor and control conditions is roughly \$11,000 in the no-explanation condition. The anchoring effect is higher still—\$31,000—when an explanation was required, but this was due largely to the presence of an award of \$1,000,000 in the control, no-explanation condition. The differences between the medians and percentiles in the anchor and control conditions provide a more reliable measure of the anchoring effect. Comparing those differences shows that both with and without explanation, the anchor shifted the awards upward throughout the full range.

Table 2c. *Damage Awards by Explanation in the Anchoring Problem*

Condition (and sample size)		Average Award (\$)	Percent \$0 Award	25th Percentile	Median	75th Percentile
No Explanation	Control (19)	60,789	21.5	500	7,500	10,000
	Anchor (20)	71,550	5.0	6500	50,000	100,000
Explanation	Control (23)	14,586	34.8	0	5,000	10,000
	Anchor (20)	46,000	10.0	5000	37,500	75,000

Because the distribution of the awards was highly skewed, assessing the results with statistical tests first required transforming the data. We took the square root of the award to produce data that more closely approximated a normal distribution and subjected the transformed data to a 2x2 ANOVA. The results showed that the anchor had a significant effect,<sup>98</sup> but neither the explanation, nor the

98.  $F(1, 78) = 8.44, p < .005$ .

interaction of explanation with the anchor, affected the awards significantly.<sup>99</sup>

2. *Framing and Fairness.* When people evaluate numeric options, they tend to code or “frame” them as either gains or losses.<sup>100</sup> Because the very same options can often be presented as either gains or losses, and because most people find losses much more unattractive than they find gains attractive, the framing of a problem can lead to impressionistic and irrational judgments.<sup>101</sup>

Framing can influence a wide range of judgments. In one study of the framing phenomenon, for example, Professors Daniel Kahneman, Jack Knetsch, and Richard Thaler gave one group of subjects a “loss” version of a problem and gave another group of subjects a “gain” version of the same problem.<sup>102</sup> In the “loss” version, the subjects learn that:

A company is making a small profit. It is located in a community experiencing a recession with substantial unemployment but no inflation. There are many workers anxious to work at the company. The company decides to decrease wages and salaries 7 percent this year.<sup>103</sup>

In the “gain” version, the subjects learn that:

99.  $F(1, 78) = 1.65, p = .20, F(1, 78) = .01, p > .90$ , respectively for the effect of the accountability manipulation and the interaction of this variable with the anchoring condition. Further analysis of the judges' gender, years of experience, and whether they make recommendations showed that these variables did not affect the judges' awards. For gender and recommendation, this was done using a 2x2x2 ANOVA on anchor condition, explanation condition, and gender; all  $F$ 's involving gender  $< 1.4, p$ 's  $> .20$ ; all  $F$ 's involving recommendation  $< 1.5, p$ 's  $> .20$ . Experience was tested with ANCOVA, using experience as a continuous variable; all  $F$ 's involving experience and its interactions  $< 1.10, p$ 's  $> .25$ . The judge's CRT score likewise had no effect. CRT score was tested with ANCOVA, using CRT score as a continuous variable; all  $F$ 's involving CRT score and its interactions  $< .15, p$ 's  $> .70$ . This last result was not entirely a surprise, as other researchers have found that scores on the CRT do not correlate with the anchoring effect. Oechssler et al., *supra* note 79, at 6 (“Furthermore, the effect of anchoring is not diminished by cognitive ability. In fact, if anything, the High CRT group seems to be more susceptible to anchoring, although this effect is not significant.”).

100. See Daniel Kahneman & Amos Tversky, *Choices, Values, and Frames*, 39 AM. PSYCHOLOGIST 341, 342 (1984) (“[T]he psychophysical analysis of outcomes should be applied to gains and losses rather than to total assets.”).

101. Daniel Kahneman, Jack L. Knetsch & Richard Thaler, *Fairness as a Constraint on Profit Seeking: Entitlements in the Market*, 76 AM. ECON. REV. 728, 731 (1986) (“These characteristics of evaluation make preferences vulnerable to framing effects, in which inconsequential variations in the presentation of a choice problem affect the decision.”).

102. *Id.*

103. *Id.*

[A company is making a small profit. It is located in a community experiencing a recession with] substantial unemployment *and inflation of 12%*. [There are many workers anxious to work at the company.] The company decides to *increase salaries only 5% this year*.<sup>104</sup>

The researchers asked both groups of subjects to evaluate the fairness of the company's conduct. In both versions of the problem, the employees end up in exactly the same position; that is, the effect of the company's wage decision is to leave employees 7 percent worse off than before. Despite the economic equivalence of the two options, however, the subjects responded quite differently to the two situations: 62 percent of subjects thought the 7 percent wage reduction in a no-inflation environment was unfair, but only 22 percent of the subjects in the 5 percent wage increase in a 12 percent inflation environment felt the same way.<sup>105</sup> Although multiple interpretations of these results are possible, the most persuasive is that the subjects reacted to the framing of the problem. The subjects interpreted the 7 percent wage reduction as a loss, but interpreted the 5 percent pay raise in a 12 percent inflation environment as a gain. As the researchers explained, "A wage cut is coded as a loss and consequently judged unfair. A nominal raise which does not compensate for inflation is more acceptable because it is coded as a gain to the employee, relative to the reference wage."<sup>106</sup>

In our previous work, we have found some evidence that framing can influence generalist judges when they evaluate settlement offers.<sup>107</sup> To explore whether framing might influence ALJs on the job, we gave the ALJs who attended the city conference a problem similar to the wage problem described above. Building on the real-world observation that companies (for example, gas stations) frequently offer cash discounts (which are viewed as a prospective gain to the consumer) but seldom impose credit card surcharges (which are viewed as a prospective loss to the consumer), we gave the ALJs a problem called the "Nonpayment Case." Some of the ALJs received a "loss" version of our problem:

---

104. *Id.* (emphasis added).

105. *Id.*

106. *Id.* at 731–32.

107. See Guthrie et al., *Judicial Mind*, *supra* note 8, at 794–97 (describing how "categorization, or 'framing,' of decision options influences the way people evaluate options and affects their willingness to incur risk" in the context of litigation).



Imagine that you are presiding in a nonpayment case filed by a landlord against a tenant whom he claims is delinquent on his rent. During the proceeding, you learn that the monthly rent is \$2,000, but the landlord imposes a \$50 per month surcharge if the tenant chooses to pay by credit card.

We gave the other ALJs a “gain” version of the same problem:

Imagine that you are presiding in a nonpayment case filed by a landlord against a tenant whom he claims is delinquent on his rent. During the proceeding, you learn that the monthly rent is \$2,100, but the landlord offers a \$50 per month discount if the tenant chooses to pay by check or cash.

We asked the ALJs in both groups to evaluate the fairness of the landlord’s conduct as “completely fair,” “acceptable,” “unfair,” or “very unfair.” In both instances, the tenant will pay the same amount—\$2,050—if she chooses the special payment option made available by the landlord.

Despite the economic equivalence of outcomes, however, the ALJs reacted quite differently to the two versions of this problem, as Table 3 shows. Of the seventeen ALJs who evaluated the fairness of the credit card surcharge, 47.1 percent found the landlord’s behavior unfair. Of the twenty-one ALJs who evaluated the fairness of the cash discount, however, a mere 4.8 percent found the landlord’s behavior unfair. The framing of the problem had a significant impact on the way the ALJs evaluated it.<sup>108</sup>

---

108. Ordered logit regression showed that the effect of framing was significant.  $z = 2.87, p < .005$ . Further analysis of the judges’ gender and years of experience, also using ordered logit regressions, showed that neither the main effect of variables nor their interaction with frame affected the judges’ ratings. For gender,  $z$ ’s  $< 1.62, p$ ’s  $> .10$ ; for experience,  $z$ ’s  $< .65, p$ ’s  $> .50$ . The judges’ CRT scores likewise did not affect the results.  $z$ ’s  $< 1.10, p$ ’s  $> .25$ . Although framing effects have been found to correlate with CRT scores in other contexts, the type of framing effect we studied here has not. Oechssler et al., *supra* note 79, at 4 (“For both items [coded as gains and losses], the High CRT group is more likely to choose the alternative that is compatible with risk neutrality.”). That is, the framing of an option as a gain or a loss influences people’s willingness to undertake risk and the framing influences how much importance people attach to gains and losses; the former, but not the latter, appears to be related to the cognitive abilities that the CRT is measuring.

Table 3. Evaluations of Fairness of Rent Payment

Condition (and sample size)	Percent who gave each evaluation (and n)			
	Completely Fair	Acceptable	Unfair	Very Unfair
Gains/Discount (21)	29% (6)	67% (14)	5% (1)	0%
Losses/Surcharge (17)	6% (1)	47% (8)	41% (7)	6% (1)

3. *Conjunction.* The so-called “extension rule”<sup>109</sup> is “perhaps the simplest and most transparent rule of probability theory.”<sup>110</sup> This rule states that “if *A* is a subset of *B*, then the probability of *A* cannot exceed that of *B*.”<sup>111</sup> For example, the probability of a terrorist act in New York City (*A*) cannot exceed the probability of a terrorist act in the United States (*B*) because the United States includes New York City (as well as many other locations that might be subjected to such an attack). Implicit in the extension rule is the “conjunction rule.”<sup>112</sup> This rule states that “the probability of *A&B* can exceed the probability of neither *A* nor *B*, since it is contained in both.”<sup>113</sup> For example, the probability of a terrorist attack in New York City carried out by Muslim extremists (*A&B*) cannot exceed the probability of a terrorist attack in New York City (*A*) or the probability of a terrorist act carried out by Muslim extremists (*B*).

The extension and conjunction rules are deductively accurate, as only a little deliberation shows. Psychologists have found repeatedly, however, that people tend to violate these rules of logic. Rather than engaging in careful deliberation, which leads to compliance with the rule, people often engage in intuitive, impressionistic thinking and thereby violate the rules. To many, it seems more likely that New York City might face a terrorist act committed by Muslim extremists than that New York City might face a terrorist attack.

---

109. Maya Bar-Hillel & Efrat Neter, *How Alike Is It Versus How Likely Is It: A Disjunction Fallacy in Probability Judgments*, 65 J. PERSONALITY & SOC. PSYCHOL. 1119, 1119 (1993); see also Amos Tversky & Daniel Kahneman, *Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment*, 90 PSYCHOL. REV. 293, 293–94 (1983) (“[P]robability theory does not determine the probabilities of uncertain events—it merely imposes constraints on the relations among them. For example, if *A* is more probable than *B*, then the complement of *A* must be less probable than the complement of *B*.”).

110. Researchers have described this as “perhaps the simplest and most transparent rule of probability theory,” Bar-Hillel & Neter, *supra* note 109, at 1130, which “even untrained and unsophisticated people accept and endorse,” *id.*

111. *Id.* at 1119.

112. *Id.*

113. *Id.*

The most famous problem of this type—the “Linda Problem”<sup>114</sup>—is instructive. In this widely administered problem, Professors Amos Tversky and Daniel Kahneman gave subjects the following information about Linda:

Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.<sup>115</sup>

The researchers asked the subjects to rank-order the likelihood of eight different statements, including these three: “Linda is active in the feminist movement”; “Linda is a bank teller”; and “Linda is a bank teller and is active in the feminist movement.”<sup>116</sup> The description made it seem as though Linda was a feminist, but not a bank teller. As a result, subjects generally reported that it was more likely that Linda was a bank teller active in the feminist movement than that she was a bank teller.<sup>117</sup> This is wrong. Under the conjunction rule, it cannot possibly be the case that it is more likely that Linda is a bank teller *and* is active in the feminist movement than that she is simply a bank teller.<sup>118</sup>

To explore whether ALJs would comply with, or violate, the conjunction rule, we gave those who attended the national conference a problem called the “Employment Case.” We asked the ALJs to imagine that they were presiding in a case involving an employment dispute between Dina El Saba, a public sector employee, and the agency for which she previously worked. The judges learn that Dina worked as an administrative assistant for a senior manager named Peter before the agency fired her. While at the agency, Dina’s

114. See Tversky & Kahneman, *supra* note 109, at 297; Amos Tversky & Daniel Kahneman, *Judgments of and by Representativeness*, in *JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES* 84, 92 (Daniel Kahneman, Paul Slovic & Amos Tversky eds., 1982).

115. Tversky & Kahneman, *supra* note 109, at 297; Tversky & Kahneman, *supra* note 114, at 92.

116. Tversky & Kahneman, *supra* note 109, at 297; Tversky & Kahneman, *supra* note 114, at 92.

117. Tversky & Kahneman, *supra* note 109, at 297; Tversky & Kahneman, *supra* note 114, at 93.

118. The Linda problem can be criticized as methodologically flawed in that people might be assuming that the single feature is actually meant to be the conjunction of “bank teller” and “not active in the feminist movement.” But Tversky and Kahneman have administered a version in which they avoid this problem by changing the second response to “Linda is a bank teller whether or not she is active in the feminist movement” and obtained similar results. Tversky & Kahneman, *supra* note 109, at 299.

employment evaluations were “average” to “above average,” so she claimed her termination must have been motivated by unlawful discrimination. The agency contends, instead, that it terminated Dina because she repeatedly violated workplace rules and norms. Among other things, she “took too many breaks during the workday and took odd days off as holidays”; “dressed in ways that made her co-workers and agency visitors feel uncomfortable, covering herself mostly in black”; acted “odd” and “aloof”; and refused to eat lunch in the presence of male coworkers.

Based solely on these facts, we asked the ALJs to rank-order the likelihood of the following four options:

- \_\_\_\_\_ The agency unlawfully discriminated against Dina based on her Islamic religious beliefs.
- \_\_\_\_\_ The agency actively recruited a diverse workforce.
- \_\_\_\_\_ The agency adhered to its internal employment policies.
- \_\_\_\_\_ The agency actively recruited a diverse workforce but also unlawfully discriminated against Dina based on her Islamic religious beliefs.

Option four (“The agency actively recruited a diverse workforce but also unlawfully discriminated against Dina based on her Islamic beliefs”) is a conjunction of option one (“The agency unlawfully discriminated against Dina based on her Islamic religious beliefs”) and option two (“The agency actively recruited a diverse workforce”). Hence, option four is, as a matter of deductive logic, less likely than either option one or option two. Nevertheless, we believed that many judges would violate the conjunction rule by identifying option four as more likely than either option one or option two (or both).

As expected, we found that the ALJs violated the conjunction rule. Rather than thinking through the problem deliberatively, which would have led them to rank-order options one and two as more likely than option four, we found the exact opposite. Of the ninety-nine ALJs who responded to this problem, eighty-four (or 84.8 percent) violated the conjunction rule in some way. These eighty-four judges committed all of the possible errors, albeit at different rates: thirty-three rated the fourth option as either equally likely as, or more likely than, both the first and second options;<sup>119</sup> thirty-six ranked the

---

119. Of these thirty-three judges, twenty ranked the fourth option as more likely than both options one and two, eight wrote in that they were all equally likely, three assigned the same

fourth option as either equally likely as, or more likely than, the first option (but not the second option);<sup>120</sup> fifteen ranked the fourth option as either equally likely as, or more likely than, the second option (but not the first option).<sup>121</sup> Thus, the problem lured most judges into committing the conjunction error, just as in the classic Linda problem.<sup>122</sup>

4. *Hindsight/Outcome Bias.* Psychologists have found that people are vulnerable to the “hindsight bias,” which is the tendency for prior outcomes to seem more predictable than was actually the case.<sup>123</sup> Once the outcome of an event is known, that outcome comes to feel inevitable or at least much more likely to have occurred than it would have seemed before it actually happened.

---

ranking to options one and two as option four, and two simply put a check mark next to option four.

120. Of these thirty-six judges, one ranked the first and fourth options as equally likely.

121. Of these fifteen judges, seven ranked the first and fourth options as equally likely.

122. Although female judges were somewhat more likely to commit the conjunction error than male judges (89 percent, or fifty out of fifty-six, versus 79 percent, or thirty-one out of thirty-nine, respectively), this difference was not significant. Fisher's exact,  $p = .24$ . More experienced judges tended to be more likely to commit the error than their younger counterparts, but this trend was also not significant. Logistic regression of committing the error on years of experience yielded a negative, but not significant, coefficient of  $-.048$ ,  $z = 1.35$ ,  $p = .18$ . The error rate was also nearly identical among judges who make recommendations as opposed to the other judges (86 percent, or thirty-six out of forty-two, versus 84 percent, or forty-two out of fifty, respectively). This was not a significant difference. Fisher's exact,  $p = 1.00$ .

Previous work on the CRT has shown that people who score high on the CRT are less likely to commit the conjunction error. Oechssler et al., *supra* note 79, at 5 (“Of our subjects in the Low CRT group, 62.6% [committed the conjunction fallacy on the ‘Linda’ problem, but t]his percentage is much lower for the High CRT group at 38.3% . . .”). We found that among the judges who scored perfectly on the CRT and answered this question, 72 percent (thirteen out of eighteen) committed the fallacy, whereas 86 percent (sixty out of seventy) of the judges who got at least one of the CRT questions wrong committed the fallacy. This difference was not, however, significant. Fisher's exact,  $p = .29$ . The percentage who committed the conjunction error broken down by exact CRT score is: zero right on CRT, 84 percent (twenty-five out of thirty); one right on CRT, 86 percent (nineteen out of twenty-two); two right on CRT, 89 percent (sixteen out of eighteen); three right on CRT, 72 percent (thirteen out of eighteen). We also ran a logistic regression of whether the judges committed the error based with CRT score as a predictor, which also showed no significant effect.  $z = .72$ ,  $p = .47$ . That we only observed a trend might be due to a somewhat small sample size with which to identify the effect.

123. See Baruch Fischhoff, *Hindsight ≠ Foresight: The Effect of Outcome Knowledge on Judgment Under Uncertainty*, 1 J. EXPERIMENTAL PSYCHOL.: HUM. PERCEPTION & PERFORMANCE 288, 288 (1975) (“Reporting an outcome's occurrence increases its perceived probability of occurrence . . .”).

Outcome information can lead not only to the hindsight bias but also to the “outcome bias.”<sup>124</sup> In contrast to the hindsight bias, which posits that outcome information can affect probability judgments, the outcome bias posits that people use outcome information to evaluate the quality of the decision made. In short, “people take outcomes into account in a way that is irrelevant to the true quality of the decision.”<sup>125</sup>

In previous work, we have found that judges are sometimes prone to the hindsight bias<sup>126</sup> and sometimes not.<sup>127</sup> To explore whether outcome information would influence ALJs, we gave the ALJs attending the city conference a problem called the “Human Rights Complaint.”<sup>128</sup> We asked the ALJs to imagine that they were presiding over a pretrial hearing in which Malcolm Jones, a fourteen-year-old African-American boy, has filed a complaint against a large toy store located in Midtown Manhattan alleging that he was harassed and detained because of his race.

The ALJs learn that Jones was in the store one day with four friends and that they were being loud, pushing and shoving, and “playing around” inside the store. Jones claims that he wandered into the video game section, played a little, and then was wrestled to the floor roughly by a guard. The guard dragged him into a back room, accused him of shoplifting, and according to Jones, subjected him to several racial slurs.

At the hearing, the guard testified that he began watching the group when they messed up the stuffed animals section of the store. He reported that he kept a close eye on Jones, in particular. He followed Jones and the other boys into the video game section and claims that he thought he saw Jones lift a video game cartridge from a countertop, prompting him to arrest Jones for shoplifting.

---

124. Jonathan Baron & John C. Hershey, *Outcome Bias in Decision Evaluation*, 54 J. PERSONALITY & SOC. PSYCHOL. 569, 570 (1988) (“At issue here is whether there is an *outcome bias* . . .”).

125. *Id.*

126. Guthrie et al., *Judicial Mind*, *supra* note 8, at 799–803.

127. See Wistrich et al., *supra* note 8, at 1313–16 (finding no statistically significant difference in judges’ conclusions based on foresight as compared to their conclusions based on hindsight).

128. Our study is similar to, and was inspired in part by, a study conducted by Professor Jonathan Casper and his colleagues. See Jonathan D. Casper, Kennette Benedict & Jo L. Perry, *Juror Decision Making, Attitudes, and the Hindsight Bias*, 13 LAW & HUM. BEHAV. 291 (1989).

The guard admitted that among the five boys who came in together, Jones was the only African-American (the others were white), but denies that this made any difference to him. The guard stated that “the kid just looked like trouble; it wasn’t because he was black.”

We informed the ALJs that they were to issue a report on this case, and we asked them to indicate whether they believed the store did or did not discriminate against Jones because of his race. Unbeknownst to the ALJs, we randomly assigned them to one of two groups, each of which received different outcome information about the alleged shoplifting. The ALJs in one group learned that “[t]he game cartridge was, in fact, found inside Jones’s coat. Jones claims he intended to pay for it, even though he did not have enough money with him to buy it.” The ALJs in the other group learned that “[t]he game cartridge was not, in fact, found on Jones. The guard claims ‘he must have ditched it somehow when he saw me coming.’” The outcome information we provided is arguably irrelevant to assessing whether the store discriminated against Jones. In both instances, there is conflicting evidence about the guard’s behavior, about Jones’s behavior, and about their respective credibility as witnesses.

To be sure, the fact that Jones was found with a stolen game cartridge in one version might undermine his credibility, and the judges might sensibly fill in any perceived gaps in the story based on their knowledge that Jones had the cartridge. But this is precisely how the hindsight and outcome biases work—people fill in the details of the story with their knowledge of the outcome.<sup>129</sup> This process might be a reasonable one if the target actor (in this case the guard) had known the details beforehand—in which case it is the outcome bias at work. But people also tend to fill in more details than were actually available to the target actor—in which case it is the hindsight bias at work. Admittedly, this problem does not perfectly distinguish between the hindsight bias and outcome bias, as a short problem will invariably leave some detail to be filled in (a videotape of the events from the guard’s point of view would have been necessary to demonstrate any difference to be the product of the hindsight bias also). But we did identify the most salient facts of this situation for the judges. And given the data on how robust and widespread the

---

129. See Jeffrey J. Rachlinski, *A Positive Psychological Theory of Judging in Hindsight*, 65 U. CHI. L. REV. 571, 584 (1998) (“[P]eople naturally integrate an outcome and the events that preceded it into a coherent story.”).

hindsight bias is,<sup>130</sup> and how it can affect situations exactly like this one,<sup>131</sup> we suggest that differences between the judgment in these two conditions would be attributable in at least some degree to improperly imputing insights to the guard that he could not reasonably be expected to have had.

We found that the ALJs responded quite differently to the two versions of this problem. Among the twenty-two judges who learned that Jones was not shoplifting, 64 percent (fourteen) of the judges determined that the guard had discriminated against Jones because of his race, whereas only 29 percent (five) of the seventeen judges who learned Jones was guilty made the same determination. This difference is significant.<sup>132</sup>

---

130. *Id.* at 581 (“[A]cross a wide variety of tasks, materials, and populations, a sizeable and consistent bias clouds judgments made in hindsight.” (footnote omitted)).

131. See Casper et al., *supra* note 128, at 300 (reporting that subjects assessing a potential civil rights violation displayed a “hindsight-bias-like process” in relying on outcome information).

132. Fisher’s exact test,  $p = .05$ .

Further analysis of gender, years of experience, and CRT score revealed some interesting trends. Male judges expressed a stronger hindsight bias than their female counterparts. Among the male judges, 71 percent (ten out of fourteen) of those who read about the innocent kid concluded that the guard had discriminated, as compared to 13 percent (one out of eight) of those who read about the guilty kid. By contrast, among the female judges, 50 percent (four out of eight) who read about the innocent kid concluded that the guard had discriminated, as compared to 44 percent (four out of nine) of the judges who read about the guilty kid. The different reaction to the knowledge of the outcome between male and female judges was marginally significant. Ordered logit of finding of discrimination on the condition, gender, and interaction of condition and gender produced a marginal effect for the interaction.  $z = 1.69$ ,  $p = .09$ .

The more experience the judges had, the less that they were influenced by the outcome, although this trend only approached significance. Ordered logit of finding of discrimination on the condition, experience, and interaction of condition and experience produced a coefficient of  $-.15$ , which was not significant.  $z = 1.55$ ,  $p = .11$ . An effect of experience here is consistent with previous research on the hindsight bias, which demonstrates smaller effects among people with greater knowledge. See Jay J.J. Chistensen-Szalansky & Cynthia Fobian Williams, *The Hindsight Bias: A Meta-Analysis*, 48 ORGANIZATIONAL BEHAV. & HUM. DECISION PROCESSES 147, 162 (1991) (“[Hindsight] effects are further reduced when the subject is familiar with the task, i.e., has expertise in or experience with the task.”).

The CRT score did not appear to influence the judges’ judgment on this task. Ordered logit of finding of discrimination on the condition, CRT score, and interaction of condition and CRT produced no significant interaction.  $z = .50$ ,  $p = .62$ . Because there is no finding that CRT score influences the hindsight bias, this was not surprising.

Because eleven of the judges identified themselves as primarily responsible for employee discipline (nine judges) or discrimination (two judges) cases, we tried to assess whether these judges behaved differently than the other judges who did not have such experience. Among the judges with no experience in this area, 71 percent (ten out of fourteen) of those who read about the innocent kid concluded that the guard had discriminated, as compared to 31 percent (four



5. *Disregarding.* People have great difficulty deliberately disregarding salient information.<sup>133</sup> In one classic illustration, researchers asked subjects participating in an elaborate experiment “not to think of a white bear.”<sup>134</sup> The subjects had great difficulty carrying out this seemingly simple task. In fact, they reported thinking about a white bear more when instructed not to do so than when they were not given such an instruction.<sup>135</sup>

In previous work, we have found that generalist judges have difficulty deliberately disregarding relevant but inadmissible evidence when making decisions, but that they can overcome their intuitive responses in some instances.<sup>136</sup> To explore whether ALJs might be similarly influenced, we gave the ALJs attending the national conference a problem called “The Alleged Rats.”<sup>137</sup>

We asked the ALJs to imagine that they were presiding over a case brought by the Department of Health against a Café for alleged health code violations involving a reported rat infestation. The ALJs learned that the city’s Health Department sent an inspector to the restaurant after neighbors complained of a rat infestation in the Café and the alleyway behind it. The inspector visited the site and found two large rat traps in the basement of the Café and an empty box of rat poison. He also took statements from a neighbor who indicated that he had complained about the rats, but that the Café’s owner had refused to do anything about it. The neighbor asserted that he had

---

out of thirteen) of those who read about the guilty kid. By contrast, among the judges with experience, 57 percent (four out of seven) who read about the innocent kid concluded that the guard had discriminated, as compared to 50 percent (one out of two) of the judges who read about the guilty kid. This suggests that experience might induce deliberative decisionmaking, but because only two experienced judges read about the guilty kid, the results cannot be interpreted as anything more than suggesting that experience aids judges’ ability to avoid the hindsight and outcome bias.

133. See Jonathan M. Golding & Debra L. Long, *There’s More to Intentional Forgetting than Directed Forgetting: An Integrative Review*, in *INTENTIONAL FORGETTING: INTERDISCIPLINARY APPROACHES* 59, 93 (Jonathan M. Golding & Colin M. MacLeod eds., 1998) (reviewing theoretical accounts of this phenomenon); Hollyn M. Johnson, *Processes of Successful Intentional Forgetting*, 116 *PSYCHOL. BULL.* 274, 274 (1994) (“[The] success of intentional forgetting depends on how one originally encoded the to-be-forgotten information . . .”).

134. Daniel M. Wegner et al., *Paradoxical Effects of Thought Suppression*, 53 *J. PERSONALITY & SOC. PSYCHOL.* 5, 6–7 (1987).

135. *Id.* at 7.

136. See Wistrich et al., *supra* note 8, at 1323–23 (concluding that although judges “do not disregard inadmissible information” in all cases, they are able to do so in some circumstances).

137. We pilot tested a version of this problem with ALJs at the city conference and made some adjustments to the context of the problem in light of recommendations from some of the judges in attendance.

seen the rats almost daily. Based on this evidence, the inspector cited the Café for a violation of the City's Health Code.

We informed the ALJs that the inspector and complaining neighbor testified to the facts above and that the neighbor also reported that he had a friend who had seen rats there. The Café owner, by contrast, denied that there was a rat problem.

The ALJs learn that per the Health Code the Health Department has the burden of proving that a violation of the Health Code occurred. They also learn that if they find that the restaurant violated the Health Code, they can impose a fine of up to \$2,000.

We divided the ALJs into three groups: a control group and two experimental groups. The ALJs in the control group received no additional information. The ALJs in the two experimental groups learned that the neighbor produced "a photo of rats on the floor of the kitchen, which he contends was taken by his friend" while his friend was visiting. The Café owner denies that the photograph depicts his kitchen, and the Café owner's lawyer moves to exclude it as improperly authenticated. We provided this additional information to the ALJs in both experimental groups, but for one of the groups we also included a copy of the photograph, showing two rats running along a floor.

We then asked the ALJs in both experimental groups how they would rule on the Café owner's motion to exclude the evidence. Finally, we compared the fines levied by the control group ALJs to the fines levied by the ALJs in the experimental groups who granted the defendant's motion to exclude the evidence. Even though these judges indicated by their rulings that they deemed the evidence irrelevant to their decisionmaking, we hypothesized that this information would nonetheless influence the fines they set in the case, particularly for those ALJs exposed to the highly salient photograph of the rats. In short, we expected that the control group ALJs would levy the lowest fines, the experimental group ALJs not exposed to the photograph would levy slightly higher fines, and the experimental group ALJs exposed to the photograph would levy the highest fines.

We found that the inadmissible evidence in this case had only limited influence on the ALJs. That is, the results reveal some trends, but no significant effects. A greater percentage of the judges who saw the photograph (85 percent, or twenty-three out of twenty-seven) than who read about it (71 percent, or twenty-two out of thirty-one)

excluded it, but this difference was not significant.<sup>138</sup> The judges who saw the photograph of the rats seemed to award higher fines. The average fine among the thirty-nine judges who did not see the photograph was \$646, as opposed to \$566 among the judges who read about (and excluded) the photograph, and \$967 among the judges who saw the photograph (and excluded it).<sup>139</sup> The awards were highly erratic, however, with many judges choosing to impose no fine (\$0), and others imposing the maximum \$2,000 fine, as Table 4 below reports. Accordingly, the apparent differences between the three versions were not significant.<sup>140</sup>

*Table 4. Fines by Condition in the Evidence Suppression Problem*

Condition (and Sample Size)	Percent Suppressing Evidence	Average Fine (\$)*	Percent Imposing \$0 fine	Percent Imposing >\$0, but <\$2000	Percent Imposing \$2000 Fine
No rats (39)	—	646	51	26	23
Read about Rats (31)	71	566	50	32	18
Saw Photo of Rats (27)	85	967	35	35	30

\* Excluding those who admitted the testimony<sup>141</sup>

6. *Egocentric Bias.* Psychologists have found that people tend to make judgments about themselves, their abilities, and their beliefs that are egocentric or self-serving.<sup>142</sup> People routinely estimate, for example, that they are well above average on characteristics that are

138. Fisher's exact test,  $p = .22$ .

139. Six judges did not provide complete responses: one judge in the condition without the evidence gave no award; one judge who read about the photo ruled it inadmissible, but gave no award; one judge (in the third condition) refused to rule or award; one judge who saw the photo ruled it admissible but gave no award; two judges who saw the photo ruled it inadmissible but gave no award.

140. A tobit regression of the award on two dummy-coded variables to reflect the three conditions did not produce a significant effects for the regression ( $p < .33$ ) or for either of the dummy codes (for exposure to the evidence versus the control condition,  $t = .22$ ,  $p = .83$ , for exposure to the photo versus the other two conditions,  $t = 1.32$ ,  $p = .19$ ).

141. Among the nine judges who read about the photo of the rats and admitted the testimony, three chose the maximum fine, three chose fines between \$0 and \$2,000 (specifically, \$400, \$500, and \$1500), and three chose no fine (\$0); the average fine was \$933. Among the four judges who saw the photo of the rats and admitted the testimony, two chose the maximum fine and two chose no fine (\$0), yielding an average of \$1,000.

142. See Michael Ross & Fiore Sicoly, *Egocentric Biases in Availability and Attribution*, 37 J. PERSONALITY & SOC. PSYCHOL. 322, 333 (1979) ("[T]he egocentric biases in availability and attribution appear to be robust and pervasive.").

important to them, such as their health,<sup>143</sup> the likely duration of their marriage,<sup>144</sup> their driving ability,<sup>145</sup> and so forth.<sup>146</sup>

In previous work, we have found that generalist judges evaluate themselves in self-serving or egocentric ways.<sup>147</sup> To explore whether ALJs might be prone to self-serving bias, we asked the ALJs attending the city conference to compare themselves to the other attendees on three dimensions: their ability to assess the credibility of a witness, their ability to avoid bias, and their ability to facilitate settlements. With regard to each, we asked the judges to place themselves into one of four quartiles: the top 25 percent, the next 25 percent, the next 25 percent, or the bottom 25 percent.

The ALJs provided self-serving interpretations of their skills, as Table 5 shows. With regard to assessing the credibility of witnesses, 83.3 percent of the ALJs placed themselves in the top half; specifically, 25 percent placed themselves in the top quartile and 58.3 percent placed themselves in the second quartile. Not a single ALJ placed herself in the bottom quartile. The ALJs were nearly as sanguine about their ability to facilitate settlements, as 86.2 percent placed themselves in the top half (30.6 percent placed themselves in the top quartile and 55.6 percent placed themselves in the second quartile). When it came to their capacity for avoiding bias in judging, a whopping 97.2 percent of the ALJs placed themselves in the top half—that is, 50 percent placed themselves in the top quartile and 47.2 percent placed themselves in the second quartile. Again, not a single ALJ placed herself in the bottom quartile. Although only 50 percent of the ALJs can be in the top half of their peers in these three

---

143. Neil D. Weinstein, *Unrealistic Optimism About Future Life Events*, 39 J. PERSONALITY & SOC. PSYCHOL. 806, 809–11 (1980).

144. Lynn A. Baker & Robert E. Emery, *When Every Relationship Is Above Average: Perceptions and Expectations of Divorce at the Time of Marriage*, 17 LAW & HUM. BEHAV. 439, 441–43 (1993).

145. Ola Svenson, *Are We All Less Risky and More Skillful than Our Fellow Drivers?*, 47 ACTA PSYCHOLOGICA 143, 145–46 (1981).

146. *But see* Don A. Moore & Deborah A. Small, *Error and Bias in Comparative Judgment: On Being Both Better and Worse than We Think We Are*, 92 J. PERSONALITY & SOC. PSYCHOL. 972, 973–74 (2007) (arguing that self-serving responses are often reasonable ones given the subjects' ignorance about how others perform on such tasks).

147. Guthrie et al., *Judicial Mind*, *supra* note 8, at 813–14.

components of their judicial behavior, most of the ALJs rated themselves this way.<sup>148</sup>

*Table 5: Self-Serving Assessments of Judicial Skills*

Skill (and sample size)	Percent in each Quartile			
	Better than 75%	Better than 50%, worse than top 25%	Better than 25%, worse than top 50%	Worse than 75%
Evaluating Credibility (36)	25	58	17	0
Facilitating Settlement (36)	31	56	8	6
Avoiding Prejudice (36)	50	47	3	0

#### D. Summary

Even though they specialize in particular areas of law, face greater accountability, and get more feedback than their generalist

148. The distribution of these results departs significantly from what one would expect if there were no bias: Evaluating credibility,  $\chi^2(3) = 26.0, p < .001$ ; facilitating settlement,  $\chi^2(3) = 24.2, p < .001$ ; avoiding prejudice,  $\chi^2(3) = 32.2, p < .001$ .

Further analysis revealed that female judges made assessments that were somewhat more self-serving than male judges as to some of their abilities. Female judges made somewhat more self-serving assessments of their abilities to assess witness credibility than male judges (76 percent, or sixteen out of twenty-one male judges, versus 93 percent, or fourteen out of fifteen female judges), but an ordered logit regression of the rating against gender revealed that this tendency was not significant.  $z = 1.42, p = .16$ . Female judges, however, rated their abilities to facilitate settlement high; whereas 76 percent (sixteen out of twenty-one) of the male judges rated themselves as better than the median judge, all fifteen female judges did so. An ordered logit regression of the rating against gender revealed that this difference was a marginally significant effect.  $z = 1.71, p = .09$ . Male and female judges made roughly equally self-serving assessments of their ability to avoid bias (the only judge to rate himself below the median was a male judge). An ordered logit regression of the rating against gender revealed no significant effect.  $z = .46, p = .64$ .

CRT score also affected the self-serving bias. Judges who scored high on the CRT made somewhat more modest assessments of their abilities to evaluate witness credibility. An ordered logit regression of the rating against CRT score revealed that this tendency was a marginally significant effect.  $z = 1.69, p = .09$ . CRT score did not correlate with the judges' assessments of their abilities to facilitate settlement. An ordered logit regression of the rating against gender revealed that this difference was a marginally significant effect.  $z = .46, p = .64$ . But judges who scored high on the CRT made more modest assessments of their abilities to avoid prejudice. An ordered logit regression of the rating against CRT score revealed that this tendency was a significant effect.  $z = 2.01, p < .05$ . These results are consistent with previous work showing that those who do well on the CRT also tend to avoid self-serving assessments. Oechssler et al., *supra* note 79, at 6 ("The fraction of subjects who judge their performance in the CRT correctly is dramatically higher for the High CRT group at 67.8% than for the Low group (9.7%).").

Years of experience did not correlate with the judges' self-serving tendencies on any of the three assessments. An ordered logit regression of the experience against each of the three evaluations revealed no significant effects: Evaluate witnesses,  $z = .45, p = .65$ ; avoid prejudice,  $z = .48, p = .63$ ; facilitate settlement,  $z = .23, p = .82$ .

colleagues, ALJs do not seem that different from the hundreds of generalist trial judges that we have studied. They performed similarly on both the CRT and on the judicial decisionmaking tasks.

These results suggest that our intuitive-override model of judging is applicable not only to generalist trial judges but also to ALJs. Both kinds of judges gravitate toward intuitive judgment on generic judgment problems and prove generally susceptible to the heuristics and biases that tend to induce intuitive and impressionistic judgments in judicial-decisionmaking problems. The data suggest that ALJs—despite their subject-matter expertise, high degree of accountability, and receipt of regular feedback—are predominantly intuitive decisionmakers, not unlike their counterparts in generalist courtrooms around the country. We found no evidence that judges who make recommendations (and hence are subject to greater feedback and accountability) react any differently to the problems we presented than their counterparts who are subject only to appeals. We also found no direct evidence that judges who are experienced in the areas that we studied were any more likely to rely on deliberation than those judges who were facing unfamiliar problems.<sup>149</sup> That said, it is worth noting that the ALJs were quite resistant to the influence of the photograph of rats in a context that was similar to the kinds of cases over which many of them preside, which is encouraging. We also found some limited evidence suggesting that expertise might facilitate better judgment. Specifically, we found that the judges with greater experience were slightly less vulnerable to the hindsight bias.<sup>150</sup>

It is also worth noting that we found that high-CRT judges were better at avoiding some of the errors in judgment than low-CRT

---

149. To be sure, our exploration of the impact of specialization on ALJ decisionmaking was imperfect. We tested two different samples of judges who spend most of their time in a wide variety of highly specialized areas. Because we could not easily anticipate what these specialties would be in advance of collecting the data and did not anticipate the degree of variation, our scenarios only infrequently and sporadically present fact patterns that match the specialty of the judges. Our scenarios for the New York judges (our first sample) included housing and discrimination issues, but we only had two judges who oversee housing cases and nine judges who address discrimination. In preparing for the study of national judges, we hoped to capitalize on the number of judges who addressed discrimination by using two scenarios involving discrimination, but the national sample included only four such judges. We also anticipated that a number of judges would identify themselves as addressing health, safety, sanitation, or environmental issues. This was true for both samples, but in the case of the New York judges, our materials were unusable, and in the national sample, we did not observe any effect that might vary by experience.

150. See *supra* note 127 and accompanying text.

judges; specifically, high-CRT judges were less egocentric<sup>151</sup> and were slightly (but not statistically significantly) less likely to commit the conjunction fallacy.<sup>152</sup> But our data do not suggest that simply hiring high-CRT judges would be a panacea for good judgment. In most contexts, judges who scored well on the CRT were just as likely to rely on misleading heuristics as those who did not.

### CONCLUSION

The role that ALJs play in the judicial process has been a matter of great debate since the inception of the ALJ position. During the second half of the twentieth century and the beginning of the twenty-first century, ALJs and their agency superiors have engaged in an ongoing tug-of-war over the issue of ALJ authority and independence. Initially, the vast majority of ALJs were embedded in one agency or another. But they have steadily become more independent. Today, roughly 60 percent of American states house ALJs in a central panel, in which they act much more like generalist trial judges and are separated from the agencies whose disputes they adjudicate.<sup>153</sup> Nevertheless, agencies generally enjoy expansive review over ALJ decisions, with no obligation to defer even to ALJ fact finding.<sup>154</sup>

In this Article, we make two contributions to this debate. First, we find no meaningful differences in the quality of ALJ decisionmaking, regardless of the degree of independence that ALJs appear to enjoy in their present positions. That is, ALJs appear to make decisions in much the same way, whether they work solely for one agency, work in a central panel jurisdiction, or enjoy some other employment arrangement. This suggests that judicial independence, though valuable for other reasons,<sup>155</sup> may not have much, if any, bearing on the quality of the justice ALJs actually dispense.

Second, our work suggests that ALJs, like generalist judges, are likely to make mostly intuitive decisions. The intuitive approach is

---

151. See *supra* note 148.

152. See *supra* note 122.

153. See *supra* note 29.

154. See *supra* note 39.

155. See, e.g., Christopher B. McNeil, *Perceptions of Fairness in State Administrative Hearings*, 92 JUDICATURE 160, 162 (2009) (reporting that citizens find some central panel adjudications more fair than within-agency adjudications).

quick, effortless, spontaneous, and often accurate.<sup>156</sup> As Professors Tversky and Kahneman observed in their early exploration of decisionmaking, intuitive thinking is “quite useful” and can often lead to accurate decisionmaking.<sup>157</sup> That said, intuition can also lead decisionmakers astray. Again, as Tversky and Kahneman observed, intuitive thinking can “lead to severe and systematic errors.”<sup>158</sup> In this Article, we found that ALJs relying on intuitive processing allowed an irrelevant anchor to influence compensatory damage awards; the framing of payment options to influence evaluations of the appropriateness of a landlord’s conduct; and the representativeness of a piece of information to influence evaluations of the likelihood of a defendant employer’s conduct. In these instances, the judges made erroneous judgments that they might have avoided by adopting a deliberative decisionmaking approach.

Combining these two observations—that ALJs make predominantly intuitive decisions regardless of their formal independence from the agencies whose disputes they adjudicate—we conclude that judicial intuition, not judicial independence, is the more vexing challenge facing executive branch judges. There are sound reasons to worry about independence, but the executive branch—like the judiciary—should also worry about minimizing, channeling, and improving judicial intuition. Perhaps more than any other reform, this will enhance justice for litigants.

---

156. See *supra* text accompanying note 18.

157. Tversky & Kahneman, *supra* note 10, at 1124.

158. *Id.*



APPENDIX A: MATERIALS USED AT THE  
NEW YORK CITY CONFERENCE

This appendix presents the text of the two hypothetical cases that we report in this paper that we used at the New York City conference, with variations noted in brackets. We also present the “skills assessment” question and the demographic questions asked at this conference.

*Nonpayment Case*

Imagine that you are presiding in a nonpayment case filed by a landlord against a tenant whom he claims is delinquent on his rent. During the proceeding, you learn that the monthly rent is [\$2,000/\$2,100], but the landlord imposes a \$50 per month [surcharge/discount] if the tenant chooses to pay by credit card. Do you think this is fair?

- Completely fair
- Acceptable
- Unfair
- Very Unfair

*Human Rights Complaint*

Suppose you are presiding over a pre-trial hearing involving the New York City Human Rights Law. Malcolm Jones, a fourteen-year old African-American boy, has filed a complaint against a large toy store in Midtown Manhattan. Jones alleges that he was harassed and detained by the store because of his race.

The hearing revealed that Jones was in the store one day with several friends. By his own admission, he and four friends were being loud, pushing and shoving, and “playing around” inside the store. He admitted that they had made a mess of the stuffed animal section. Jones claimed they moved on to play some of the video games that the store makes available for customers. Jones claims that he wandered around in the section, played a little, and then was wrestled to the floor roughly and handcuffed by a guard. The guard dragged him into a back room and accused him of shoplifting. Jones and two of his friends contend that the guard used several racial slurs.

At the hearing, a guard testified that he began watching “that little pack” when they messed up the stuffed animals. “I kept a close

eye on that one especially”, as he pointed to Jones. He stated that he followed the boys into the video game section to keep an eye on them. He stated that a small game cartridge had been placed temporarily by a clerk onto a countertop, and that “that punk spotted it right away.” The guard claimed that he saw Jones ease over to the countertop, and when he walked away, “the cartridge was gone.” The guard then arrested Jones for shoplifting.

The guard admitted that among the five boys who came in together, Jones was the only African-American (the others were white), but denies that this made any difference to him. The guard stated, “that kid looked like trouble; it wasn’t because he was black.”

The game cartridge was, in fact, [found inside Jones’ coat/not found on Jones]. [Jones claims he intended to pay for it, even though he did not have enough money with him to buy it./ The guard claims “he must have ditched it somehow when he saw me coming.”]

You must issue a report in which you conclude whether the store likely discriminated against Jones because of his race, or not. Given these facts, what is your opinion?

- Yes, Jones was likely discriminated against because of his race
- No, Jones was likely not discriminated against because of his race

### *Skills Self-Assessment*

Relative to the other judges attending this conference, how would you rate yourself on the following:

Assessing the credibility of a witness

- In the highest quartile (meaning that you are more skilled at this than 75% of the judges attending this conference)
- In the second highest quartile (meaning that you are more skilled at this than 50% of the judges in this room, but less skilled than 25% of the judges attending this conference)
- In the second lowest quartile (meaning that you are more skilled at this than 25% of the judges in this room, but less skilled than 50% of the judges attending this conference)
- In the lowest quartile (meaning that you are less skilled at this than 75% of the judges attending this conference)

[Note that we also used the same 4 categories for the three skills listed below:]

Avoiding racial bias in making decisions

Facilitating settlements

*Demographic Information*

Please identify your gender:

male

female

For how many years have you served as a judge? \_\_\_\_ years

What type of cases do you normally hear? \_\_\_\_\_

Which of the two major political parties in the United States most closely matches your own political beliefs?

The Republican Party

The Democratic Party

## APPENDIX B: MATERIALS USED AT THE NATIONAL CONFERENCE

This appendix presents the text of the three hypothetical cases that we report in this paper that we used at the New York City conference, with variations noted in brackets. We used the same demographic questions as in the New York City Conference.

*Employment Case*

Imagine that you are presiding in a case involving an employment dispute between Dina El Saba, a public sector employee, and the agency for which she previously worked. Dina was an administrative assistant for a senior manager named Peter before the agency terminated her employment. At the agency, her employment evaluations were all “average” to “above-average,” so she contends her termination was motivated by unlawful discrimination. Peter concedes that Dina’s performance evaluations were as she claims, but he reports that the agency terminated her for repeatedly violating workplace rules and norms. Among other things, Dina took too many breaks during the workday and took odd days off as holidays. He also claims she dressed in ways that made her coworkers and agency visitors feel uncomfortable, covering herself mostly in black. He also contends that she acted “odd” and “aloof”, refusing to eat lunch while male coworkers were present in the break room.

Based solely on these facts, how likely is it that: (Please rank these in order of likelihood, where “1” is the most likely, “2” is the second-most likely, “3” is the third-most likely, and “4” is the least likely.)

- \_\_\_\_\_ The agency unlawfully discriminated against Dina based on her Islamic religious beliefs.
- \_\_\_\_\_ The agency actively recruited a diverse workforce.
- \_\_\_\_\_ The agency adhered to its internal employment policies.
- \_\_\_\_\_ The agency actively recruited a diverse workforce but also unlawfully discriminated against Dina based on her Islamic religious beliefs .

*Veronica v. AAC*

Imagine that you are presiding over an employment discrimination case brought by the City Commission on Human Rights (“Commission”) on behalf of Veronica Sanchez, a Mexican-

American who legally immigrated to the U.S., against Administrative Assistants on Call, LLC. (“AAC”). The Commission alleges that AAC violated the Administrative Code by terminating Veronica based on her “actual or perceived” race, color, national origin, and alienage or citizenship status.

AAC enters into contracts with small businesses to provide off-site secretarial and administrative support services. Until six months ago, Veronica was one of AAC’s top employees. Six months ago, however, AAC hired a new manager, and according to Veronica, he immediately gave her only undesirable projects. Veronica testified that he used racially offensive terms in her presence; told her on a couple of occasions to “go back to Mexico”; and accused her of taking jobs from “real Americans.” On one occasion, he made comments like these in front of one of Veronica’s former coworkers, who corroborated Veronica’s testimony, and in front of Veronica’s daughter, who was visiting her at work that day. Veronica complained to the manager, who then fired her.

AAC does not deny Veronica’s allegations, but AAC contests the Commission’s request for damages on Veronica’s behalf. AAC’s owner testified that the company eventually fired the manager and noted that Veronica found comparable employment shortly after her termination, and she therefore does not even make a claim for lost wages.

The Commission seeks compensatory damages on Veronica’s behalf. Under the Administrative Code, you are authorized to award compensatory damages for “mental anguish,” and controlling case law indicates that an award is justified where “there is credible evidence of some degree of emotional distress.” Here, Veronica testified credibly that she has suffered from anxiety, sleeplessness, and bad dreams. She also mentioned as an aside that she recently saw a case similar to hers on a “court television show” where the plaintiff received a [\$415,300] compensatory damage award for mental anguish.

Commission practice gives you broad discretion, [and experience indicates that review of your award by the Commission’s appellate panel is extremely unlikely/ but experience indicates that review of your award and your explanation by the Commission’s appellate panel is very likely.]

Based solely on these facts, what compensatory damages would you award Veronica for mental anguish?

\$ \_\_\_\_\_

Please briefly explain the basis for your damage award:

---

---

---

---

---

### *The Restaurant*

Imagine that you are presiding over a case brought by the Department of Health against a restaurant called Café Des Artists for an alleged Health Code violation.

The source of the complaint is a citation against the Café for an alleged rat problem. The City Health Department sent an inspector to the restaurant after neighbors complained of a rat infestation in the Café and the alleyway behind it. The inspector visited the site and found two large rat traps in the basement of the Café and an empty box of rat poison. He also took statements from a neighbor who indicated that he had complained about the rats, but that the Café's owner has refused to do anything about it. The neighbor asserted that he had seen the rats almost daily, mostly "coming out of the kitchen". Based on this evidence, the inspector cited the Café for a violation of the City's Health Code.

Both the neighbor and the inspector testified at the hearing. The inspector restated the facts above. The neighbor added that he has a friend who also saw rats when he visited from out of town, but this friend was not available to testify. [He produced a photo of rats on the floor of the kitchen (*see below*), which he contends was taken by his friend through an open door of the restaurant's kitchen. He claims the friend sent him the photo by e-mail.]

At the hearing, the Café's owner denied that there are rats in his restaurant. He claims he has seen no rats in the restaurant and that he has received no complaints from the neighbors. He asserts that the traps and empty box of poison must have been left there by a previous owner before he took over the restaurant a year ago. [He denies that the photo was taken in the restaurant. At the hearing his lawyer objected to the use of the photo as lacking sufficient authentication and moved that it be excluded.]

ould you rule on the motion to exclude the photograph?  
y the motion and admit the photograph into evidence  
nt the motion and exclude the photograph from  
evidence]

Health Codes specifies that the Health Department has the  
demonstrating that a violation occurred. If you find that a  
violation occurred, then you may levy a fine of up to \$2,000.

Based solely on the evidence admitted at the hearing, what fine,  
if any, would you impose on the Café owner?

---

*The Photo:*

