Can the Dark Arts of the Dismal Science Shed Light on the Empirical Reality of Civil Procedure?

Jonah B. Gelbach

University of Pennsylvania Carey Law School

Follow this and additional works at: https://scholarship.law.upenn.edu/faculty_scholarship

Part of the Behavioral Economics Commons, Civil Procedure Commons, Courts Commons, Law and Economics Commons, Litigation Commons, Policy Design, Analysis, and Evaluation Commons, Political Economy Commons, Public Policy Commons, and the Social Statistics Commons

Repository Citation

https://scholarship.law.upenn.edu/faculty_scholarship/1046

This Article is brought to you for free and open access by Penn Law: Legal Scholarship Repository. It has been accepted for inclusion in Faculty Scholarship at Penn Law by an authorized administrator of Penn Law: Legal Scholarship Repository. For more information, please contact PenncollIR@law.upenn.edu.
CAN THE DARK ARTS OF THE DISMAL SCIENCE SHED LIGHT ON THE EMPIRICAL REALITY OF CIVIL PROCEDURE?

Jonah B. Gelbach*

ABSTRACT

Litigation involves human beings, who are likely to be motivated to pursue their interests as they understand them. Empirical civil procedure researchers must take this fact seriously if we are to adequately characterize the effects of policy changes. To make this point concrete, I first step outside the realm of civil procedure and illustrate the importance of accounting for human agency in empirical research. I use the canonical problem of demand estimation in economics to show how what I call the “urn approach” to empirical work fails to uncover important empirical relationships by disregarding behavioral aspects of human action.

I then show how these concerns permeate a prominent empirical issue in contemporary civil procedure debates: the changes in pleading policy wrought by Bell Atlantic, Corp. v. Twombly and Ashcroft v. Iqbal. Revisiting my own earlier work, I embed the question of how changes in the pleading standard will affect case outcomes in a broad behavioral framework that takes parties’ agency seriously. In the process, I address recent critiques, both of the very idea of using behavioral frameworks to understand civil litigation policy changes, and of certain aspects of my use of real-world litigation data collected by the Federal Judicial Center. As I show, these criticisms are straightforwardly refuted on the merits.

The alternative to taking seriously the behavioral context created by the civil justice system—what has occurred so far in too much of the debate over Twombly and Iqbal—is, as one critic of early 20th-century empirical research by legal scholars once put it, “a mindless amassing of statistics without reference to any guiding theory whatsoever.” To do better, we will need to take behavior seriously in studying civil litigation.

INTRODUCTION ..........................................................225

I. TWOMBLY AND IQBAL: WHY THEY MATTER, AND WHY EMPIRICAL RESEARCH CONCERNING THEIR EFFECTS PROVIDES A GOOD LENS ..........................................................229

* Associate Professor, University of Pennsylvania Law School. I thank Steve Burbank, Joe Cecil, David Freeman Engstrom, Joshua Fischman, Jacob Goldin, William Hubbard, Jon Klick, Bruce Kobayashi, Zach Liscow, Alex Reinert, David Schleicher, Judge Anthony Scirica, Cathie Struve, Margaret Williams, and Tobias Barrington Wolff for helpful discussions.

223
A. The Grant Rate-Based Empirical Literature on Twombly

STANFORD JOURNAL OF COMPLEX LITIGATION [Vol. 2:2

B. A Brief Description of Locking the Doors

II. BEHAVIORALLY INDIFFERENT EMPIRICS HAS URNED ITS

RETIREMENT

A. Behavioral Challenges Illustrated: Estimating a Demand

Function

B. The Absence of Party Selection as a Behavioral Restriction

III. A CASE STUDY: THE LOCKING THE DOORS BEHAVIORAL

FRAMEWORK AND EMPIRICAL IMPLEMENTATION

A. Potential Outcomes

B. Using an Economic Model as Behavioral Framework

Eliminates the Possibility of Some Dispute Types, But Not

Enough to Make Grant Rate Differences Informative

C. Mapping Potential Outcomes into a Measure of Negatively

Affected Cases

D. Empirical Implementation

E. Empirical Evidence: Estimates of Alternative Lower Bounds

F. Robustness of the Empirical Lower Bound Results

1. Setting fire to straw men: Priest & Klein and the 50-

percent hypothesis

2. Are there major problems with the model’s behavioral

assumptions?

a. Other Rule 12(b) motions

b. Leave to amend

c. Multiple claims

d. Multiple parties

3. The Proper Unit of Analysis: Claims, Cases, or

Plaintiffs?

a. The “wrong denominator” problem

b. A feasible alternative version of Engstrom’s

approach yields results that buttress the qualitative

conclusions in Locking the Doors

G. Potential Confounding Factors, and a Previously Unnoted

Data Issue

1. Ruling Out Confounding Factors

2. “Type Z” Disputes

CONCLUSION

APPENDIX: A ROUGH GUIDE TO THE EMPIRICAL IMPORTANCE OF TYPE

Z DISPUTES
INTRODUCTION

Normatively relevant—and controverted—empirical claims abound in the civil procedure arena:

• “[T]he success of judicial supervision in checking discovery abuse has been on the modest side.”

• “[A]mendments to the Federal Rules and changes in various judicial practices have been designed, for more than a quarter century, to contain or control discovery and enhance the power of judges to manage cases throughout the pretrial process. . . . [S]ome believe [these developments have] enabled defense interests to employ the procedural system to avoid, or at least delay, reaching an adjudication of a dispute’s merits.”

• “All fair observers acknowledge the skyrocketing cost of discovery.”

• “[E]mpirical research on discovery conducted over thirty years has not demonstrated that it has been a problem in more than a small slice of litigation.”

Yet empirical research in civil procedure is vulnerable to a charge directed at the field of empirical legal studies more generally, that it sometimes involves “a mindless amassing of statistics without reference to any guiding theory whatsoever.” This quotation is drawn from an important recent paper by Joshua B. Fischman, which advocates *Reuniting ‘Is’ and ‘Ought’ in Empirical Legal Scholarship.* As Fischman sees things, “[t]he fundamental problem is that empirical legal methodology lacks frameworks for connecting empirical findings with normative conclusions.” Concentrating his fire on studies of judicial citation counts, reversal rates, and interjudge disparities, Fischman argues that such studies “conflate the measureable with the good, justifying policy proposals on the basis of the measureable objects.” In the present Article, I both apply and build on Fischman’s critique with respect to one important strand of research in civil procedure—pleading standards and the effects of *Bell Atl. Corp. v. Twombly,* 550 U.S. 544, 559 (2007).

---

4. *Id.* at 151 (rebuttal of Burbank).
7. *Id.* at 154.
8. *Id.* at 121.
I focus my discussion through the pleading standards lens for three reasons. First, the Supreme Court’s decisions in *Twombly* and *Iqbal* touched off a firestorm of controversy, much of it centered on empirical questions such as whether *Twombly* and *Iqbal* closed off access to court by making it harder to get to discovery. Whether for better or worse, that was the point of *Twombly*, at least with respect to antitrust cases. Policy makers and stakeholders in the civil justice system care about the effects of *Twombly* and *Iqbal*—and they should. Second, pleading standard changes provide an excellent field for understanding how researchers might meet Fischman’s entreaty to “be more explicit about how they are combining objective findings with contestable assumptions in order to reach normative conclusions.” Contemporary applied economics is all about making such combinations in systematic ways, and there are some straightforward ways in which the study of *Twombly* and *Iqbal*’s empirical effects could be improved using methodological insights from this field.

Third, most empirical studies of whether *Twombly* and *Iqbal* have reduced plaintiffs’ access to court have fallen into just the trap that Fischman describes. As I shall discuss, these studies focus primarily on how one or another measure of the Rule 12(b)(6) grant rate has changed. Unfortunately, measuring this outcome does not answer the question authors seem to think it does—in Fischman’s terms, the measureable object is disconnected from what even those measuring it see as the policy-relevant one. Much ink has spilled in debates over such details as:

- whether it is acceptable to use electronic data bases such as Westlaw, or whether cases studied must be drawn directly from administrative sources (such as PACER-hosted docket reports or a data base to which only FJC authors have had access);13

---

11. *Id.* at 559 (“It is no answer to say that a claim just shy of a plausible entitlement to relief can, if groundless, be weeded out early in the discovery process through careful case management, . . . given the common lament that the success of judicial supervision in checking discovery abuse has been on the modest side. . . . the threat of discovery expense will push cost-conscious defendants to settle even anemic cases before reaching those proceedings.”) (citations and quotation marks omitted).
• whether pro se cases should be included; and whether pro se cases should be included;

• and how one should account for the possibility that Rule 12(b)(6) motions might be granted with leave to amend.

But those engaged in these arguments about the trees have sped right past the forest. As I have previously argued, rational parties can be expected to change their litigation behavior in response to perceived changes in the pleading standard. One consequence is that changes in the grant rate by themselves generally tell us nothing at all about how judicial behavior has changed as a result of Twombly and Iqbal.

Another important consequence of party selection is that changes in judicial behavior aren’t the only factor of normative interest, because party selection can have direct effects on litigation outcomes. For example, an elevated pleading standard might cause a defendant to file a Rule 12(b)(6) motion that she wouldn’t have filed under the more liberal Conley v. Gibson standard. If such a motion is granted, the plaintiff will lose at the pleading stage under Twombly/Iqbal, but not under Conley. That is an effect we should want to measure in order to understand the full normatively relevant scope of Twombly and Iqbal’s effects.

(2012) [hereinafter Moore, Updated Quantitative Study]. Beyond those stated in the text herein, I take no positions on either these studies or Cecil’s criticisms of them. See also, e.g., Cecil, Waves, supra at 3-4 (acknowledging that the FJC authors missed some motions and promising to “reanalyze the data to determine what effect this has on our original findings”); Cecil, Waves, supra at 27-31 (arguing that electronic databases likely cause biases); David Freeman Engstrom, The Twombly Puzzle and Empirical Study of Civil Procedure, 65 STAN. L. REV. 1203, 1214-17 (2013) [hereinafter Engstrom, Twombly Puzzle]. Engstrom considers and critiques an even larger set of Twombly/Iqbal-related empirical studies than does Cecil. See, e.g., Engstrom, Twombly Puzzle, supra at 1245 (considering twelve studies, including Locking the Doors). I take no position on Engstrom’s arguments except insofar as the text herein states. See also Hoffman, supra at 35 (showing that the database search conducted by the authors of the FJC Initial Report missed at least some Rule 12(b)(6) motions); Moore, Updated Quantitative Study, supra at 608 (“[D]istrict court orders ruling on 12(b)(6) motions in Westlaw are fairly representative of the universe of all such district court orders”) (emphasis in original)).


15. See, e.g., Cecil, Waves, supra note 13, at 23; Hoffman, supra note 13, at 32-34; Moore, Updated Quantitative Study, infra note 13, at 639-40.

16. See, e.g., Engstrom, Twombly Puzzle, supra note 13, at 1221-22 (characterizing as “strange research design choices” authors’ failure to “adequately distinguish between 12(b)(6) grants with and without leave to amend”).


18. See Part III, infra.

Thus, those who base policy suggestions—whether to reverse or stay the course—on changes in the grant rate perfectly illustrate Fischman’s concern about “confl[ating] the measureable with the good, justifying policy proposals on the basis of the measureable objects.”  

Ironically, despite the widely recognized normative relevance of changes in pleading policy, the overwhelming majority of studies that seek to measure *Twombly* and *Iqbal*’s effects fail Barry Friedman’s well-taken admonition: “Normative bite ought to define the problem, not be an afterthought. Falsifiable hypotheses should be about something of consequence.” In the empirical literature on *Twombly* and *Iqbal*, the grant rate has been the object about which falsifiable hypotheses have been constructed. Yet authors have failed to take seriously the question of whether changes in the grant rate tell us anything about the issues of normative concern. As to this point, Fischman could not be more on point when he writes that “research should focus on what is important, not what is easily measureable.”

The fundamental problem with such civil procedure empiricism, then, is not that it lacks normative motivation. Rather, such empiricism fails to adequately represent the behavioral determinants of normatively relevant positive objects. If parties change their behavior when procedural rules change, then understanding the empirical effects of rules changes requires treating party behavior as a fundamental aspect of the conceptual framework that guides empirical work. In short, the facts of litigation are behaviorally and socially generated, and such facts cannot be usefully studied by pretending they simply involve the classical statistics teaching tool of drawing balls from an urn.

A core objective of this Article is to show how approaches widely used by applied economists can help increase the relevance—both positive and normative—of empirical civil procedure research. The point here is not to make civil procedure an object of methodological economics imperialism. Rather, it is to take to heart the counsel provided by Fischman and by Friedman. Empirical questions in civil procedure are too important to study as if human behavior weren’t involved, just as they are too important to study with data or statistical methods that obscure rather than illuminating.

In Part I, I provide a brief refresher on the Supreme Court’s pleading jurisprudence, as well as on the empirical literature that has attempted to study *Twombly* and *Iqbal*’s effects. In Part II, I characterize and criticize the approaches widely used by applied economists can help increase the relevance—both positive and normative—of empirical civil procedure research. The point here is not to make civil procedure an object of methodological economics imperialism. Rather, it is to take to heart the counsel provided by Fischman and by Friedman.

---

21. Barry Friedman, *Taking Law Seriously*, 4 PERSP. ON POL. 261, 263 (2006). The irony is that in writing these sentences, Friedman was actually criticizing positive scholarship concerning the law for failing to take seriously law’s internal norms and concerns, a deficiency that hardly can be pinned to *Twombly/Iqbal* researchers.
23. See Part II, infra, for the relevance of this analogy.
24. At a minimum, we should avoid winding up feeling like Karl Llewellyn, who described his response to early legal empiricism thusly: “I read all the results, but I never dug out what most of the counting was good for.” Karl N. Llewellyn, *On What Makes Legal Research Worth While*, 8 J. LEGAL EDUC. 399, 403 (1956).
theoretical, behavior-neutral approach to empirical research that most such studies have taken. I begin with deliberately far-afield examples—sampling balls from an urn and the estimation of a demand function in economics. These examples allow me to abstract from issues that are contested in the civil procedure arena, while still illustrating how a failure to take social and behavioral facts seriously can doom the usefulness of empirical work from the get-go. I then circle back to the challenge of measuring Twombly and Iqbal’s effects, and the extant literature’s general failure to take changes in party behavior seriously.

In Part III, I adopt the conceptual framework used in my earlier work, Locking the Doors to Discovery. That paper takes party selection behavior seriously and develops a measure that is empirically informative as to Twombly and Iqbal’s effects. Finally, I conclude.

I. TWOMBLY AND IQBAL: WHY THEY MATTER, AND WHY EMPIRICAL RESEARCH CONCERNING THEIR EFFECTS PROVIDES A GOOD LENS

Pleading, it is said, is the gateway to the courts. And since Twombly and Iqbal, it has certainly become a controversial gateway. Twombly and Iqbal (in)famously replaced Conley v. Gibson’s “no set of facts” standard for determining when a district court judge should grant a defendant’s Rule 12(b)(6) motion for failure to state a claim. In place of this logical-possibility test for a complaint’s legal sufficiency, Twombly and Iqbal have substituted the much-debated plausibility standard. Under the plausibility standard, a judge is meant to examine the complaint’s non-conclusory allegations to “determine whether they plausibly give rise to an entitlement to relief,” and this inquiry is supposed to be controlled by judges’ “judicial experience and common sense.”

Changes in the pleading standard could easily work substantial changes on access to court, the extent of frivolous litigation, and the mix of disputes that are litigated. The Supreme Court’s decisions in Twombly and Iqbal spawned

25. Gelbach, Locking the Doors, supra note 17, at 2270.
27. Conley v. Gibson, 355 U.S. 41, 45-46 (1957) (“In appraising the sufficiency of the complaint we follow, of course, the accepted rule that a complaint should not be dismissed for failure to state a claim unless it appears beyond doubt that the plaintiff can prove no set of facts in support of his claim which would entitle him to relief.”).
28. Ashcroft v. Iqbal, 556 U.S. 662, 679 (2009) (“In keeping with these principles a court considering a motion to dismiss can choose to begin by identifying pleadings that, because they are no more than conclusions, are not entitled to the assumption of truth. While legal conclusions can provide the framework of a complaint, they must be supported by factual allegations. When there are well-pleaded factual allegations, a court should assume their veracity and then determine whether they plausibly give rise to an entitlement to relief.”).
29. Id.
widespread concern and attempts at congressional override, as well as challenges to the premise of the plausibility-pleading regime. The importance of Twombly/Iqbal empirics is the genesis of two reports issued by authors affiliated with the Federal Judicial Center (“FJC”). The FJC “assess changes in motions to dismiss and decisions on such mo-


tions over time in broad categories of civil cases.”

Given the prominent role that pleading plays in litigation, it is no surprise that Twombly and Iqbal have been cited gajillions of times—in more than 90,000 and 70,000 cases, respectively, as of this writing. Moreover, a forest of trees surely must have been hacked down to print all the scholarly articles written about them in the last several years, including many that attempt to measure empirical effects of Twombly and Iqbal using the change in the share of Rule 12(b)(6) motions that are granted (according to one measure or another of “grant”).

And the Supreme Court’s interest in pleading is not over. For example, one of the two questions presented in Wood v. Moss, argued at the Court on March 26, 2014, directly concerns the factual sufficiency of a complaint in a constitutional civil rights case. Finally, discussions related to Twombly and Iqbal plug into a broader debate about the extent and nature of litigation, discovery costs, and judicial policy. As just one example, consider the August 2013 promulgation by the Advisory Committee on Rules of Civil Procedure of a Preliminary Draft of Proposed Amendments to the Federal Rules of Civil Procedure. Included in these amendments are various changes to the discovery rules, including ones related to proportionality that are directed at reducing the cost of discovery—a major issue in both the Supreme Court’s opinions in Twombly and Iqbal and the debates that have followed them. Over 2,000 public comments on the proposed amendments were submitted, and many of these raise the issue of empirical evidence.

34. FJC Initial Report, supra note 33, at vii; FJC Updated Report, supra note 33, at 1.
35. See infra note 40.
A. The Grant Rate-Based Empirical Literature on Twombly and Iqbal
232 STANFORD JOURNAL OF COMPLEX LITIGATION (Vol. 2:2)

There is by now a sizable number of papers that have sought to measure the effects of Twombly and Iqbal.40 The most frequent object of measurement has been the change in one measure or another of the outcome of Rule 12(b)(6) motions. Sometimes the Rule 12(b)(6) outcome has been measured such that a motion is coded as granted if it is granted as to one or more claims,41 sometimes only grants that eliminate at least one plaintiff are counted,42 and sometimes only case terminations count.43 Studies also differ as to how much attention they pay to whether motions were granted without leave to amend or with prejudice.

Studies have focused on a wide array of case types, though constitutional civil rights and discrimination-related cases have received special attention.44 There has been substantial debate over whether prisoner and pro se cases should be included or excluded.45 Some studies present only raw grant rate data,46 while others report estimates from multivariate models intended to control for variation in factors other than Twombly and Iqbal.47 Different studies


41. See, e.g., Moore, Updated Quantitative Study, supra note 13; Dodson, supra note 40; FJC INITIAL REPORT, supra note 33, at 13-15; FJC UPDATED REPORT, supra note 33, at 7-8, Tables A-1 and A-2.

42. See, e.g., FJC INITIAL REPORT, supra note 33, at 17-19; FJC UPDATED REPORT, supra note 33, at 7-8, Tables A-1 and A-2.

43. See, e.g., Hubbard, Testing for Procedural Change, supra note 40; FJC INITIAL REPORT, supra note 33, at 16; FJC UPDATED REPORT, supra note 33, at 10, Table A-4.

44. See, e.g., Brescia, supra note 40; Moore, Updated Quantitative Study, supra note 40.

45. See studies cited supra notes 14 and 15.

46. See, e.g., Seiner, Pleading Disability, supra note 40; Seiner, Trouble with Twombly, supra note 40; Brescia, supra note 40.

47. See, e.g., FJC INITIAL REPORT, supra note 33; FJC UPDATED REPORT, supra note
have reported widely varying changes in their measures of the grant rate, both within and across case type categories. While authors generally have not been entirely clear about the point of studying these changes in grant rates, one relatively explicit declaration comes from William Hubbard, who indicates his focus is in “[q]uantifying change in legal standards—in the sense of change in the propensity of judges to decide cases a certain way.” This characterization suggests that Hubbard has in mind what I call the “judicial behavior effect.” By this term, I mean the difference in the probability that a movant would win on a Rule 12(b)(6) motion that is but-for caused by Twombly and Iqbal (however one measures winning). As I shall discuss in detail in Parts III.B and III.C, infra, the critical problem is the potential presence of various types of party selection effects. Judicial behavior effects seem to be the focus not only of Hubbard’s work, but of other studies as well.

As I discuss in Part III.B, infra, these studies’ estimates of the change in grant rates generally would represent judicial behavior effects only under the implausible assumption that there are no changes in party behavior in response to Twombly and Iqbal. Thus, none of these studies should be viewed as providing a credible estimate of the judicial behavior effect of Twombly and Iqbal on any identifiable set of cases.

Moreover, even if some measure of judicial behavior effects could be identified empirically, the possibility of party selection effects requires us to know

33; Moore, Tao of Pleading, supra note 13; Hubbard, Theory of Pleading, supra note 40; Hubbard, Testing for Procedural Change, supra note 40; Moore, Updated Quantitative Study, supra note 13.

48. See, e.g., Quintanilla, supra note 40, at 5 (reporting an increase of thirty-four percentage points in the rate at which black plaintiffs’ claims of race discrimination are dismissed in employment discrimination cases); FJC INITIAL REPORT, supra note 33, at 14, Table 4 (reporting increase of 3.2 percentage points in the share of employment discrimination cases granting some or all relief to Rule 12(b)(6) movants); FJC UPDATED REPORT, supra note 33, at 7, Table A-1 (reporting change of only 0.2 percentage points for employment discrimination cases after following up cases with amended complaints); FJC UPDATED REPORT, supra note 33, at 7, Table A-1 (finding increase of 6.2 percentage points in share of cases in which Rule 12(b)(6) movant prevailed on some or all claims, after accounting for any amended complaints, with much of this difference apparently driven by increase of forty-two percentage points among financial instruments cases, which might have been due in part to the financial and housing crisis rather than to changes in the pleading standard).

49. Hubbard, Testing for Procedural Change, supra note 40, at 35.

50. See Part III.A, infra.

51. As I discuss in Parts III.B and III.C, infra, Hubbard’s studies are not immune to this problem, despite his claims to the contrary. See Hubbard, Testing for Procedural Change, supra note 40, at 38 (“This paper addresses both the selection of disputes into lawsuits and the selection of lawsuits into adjudication.”); Hubbard, Theory of Litigation, supra note 40, at 16 (“By limiting my analysis to cases filed before the decision, I was able to control for selection effects.”).

52. See Cecil, Waves, supra note 13, at 46 (in characterizing the FJC authors’ updated report and the outcomes-oriented parts of their initial report, for example, Cecil states that their purpose was “to assess the reaction of the courts” to Twombly and Iqbal).
more to meet Fischman’s exhortation to “prioritize normative questions.”

From the perspective of a given party in a given case, the probability that a Rule 12(b)(6) motion would be successful—however that is measured—might be the only question of interest. But from the normative perspective of legal policy—of the design of the civil justice system—that is surely not the only question of interest, because changes in party behavior can themselves affect parties’ welfare. To even begin to understand the normative consequences of Twombly and Iqbal, we must take account of these effects, too.

That brings me back to the FJC authors’ initial report, which is the only study to date that provides direct evidence on Twombly/Iqbal-induced changes in the filing of Rule 12(b)(6) motions. As I discuss infra, pre-Twombly and post-Iqbal information on the change in filings can be used with grant rate data to at least partially illuminate the effects of Twombly and Iqbal—even without assuming there is no party selection.

B. A Brief Description of Locking the Doors

In my recent work on this topic, Locking the Doors to Discovery? (“Locking the Doors”), I sought to recast the discussion of Twombly and Iqbal’s empirical effects away from the overly simplistic metric of grant rate changes. I argued in Locking the Doors that we should expect a perceived change in pleading standards to induce changes in parties’ lawsuit-filing, Rule 12(b)(6) motion-filing, and settlement behavior. Without unpalatable and empirically untestable assumptions, such party selection effects render even the direction of change in the grant rate uninformative about how judges have responded to the change. However counter-intuitively, a drop in the grant rate is consistent with a pro-defendant judicial behavior change. And an increase in the grant rate is consistent with the absence of any change in judicial behavior. What’s more, party selection effects don’t just obscure changes in judicial behavior—they have direct effects on case outcomes, too.

53. Fischman, supra note 5, at 154.
54. Note that even the broader accounting of the effects of Twombly and Iqbal proposed here does not exhaust the normatively relevant domain of questions related to pleading changes. For one thing, it cannot measure all the negatively affected plaintiffs. And even if many plaintiffs are negatively affected by Twombly and Iqbal, in the sense I define infra, Twombly and Iqbal might be socially beneficial on net: if the negatively affected cases have low enough merit, then it is a good thing for them not to get to discovery. The question of how Twombly and Iqbal have affected case quality is the subject of my work in process, which will use case outcomes after the answer and Rule 12(b)(6) stage to try to measure case quality effects.
55. FJC Initial Report, supra note 33, at 8-12.
56. Gelbach, Locking the Doors, supra note 17.
57. For example, consider a case in which a Rule 12(b)(6) motion would be filed, and then granted, under Twombly/Iqbal, but in which that motion would not be filed in the first place pre-Twombly. Obviously the plaintiff would be worse off—and the defendant better off—if this case were litigated under Twombly/Iqbal than if it were litigated under Conley. Yet I have said nothing about what would have happened if the defendant were to file a Rule
In Locking the Doors I showed how existing data could be used to place a lower bound on the magnitude of Twombly and Iqbal’s negative effects on those plaintiffs who actually face Rule 12(b)(6) motions under Twombly and Iqbal. Using data published in the two FJC reports, I showed that plaintiffs were negatively affected—a term I define precisely in Locking the Doors and revisit below—in a non-trivial share of such cases.

Two relatively detailed critiques of Locking the Doors have since appeared. One is Joe Cecil’s Of Waves and Water (“Waves”). Cecil is the lead author of the two FJC-released reports, and while his critique is unpublished at present, his views, arguments, and empirical claims are very important to civil litigation policy. As he is a senior researcher at the FJC who frequently provides detailed empirical analysis to members of the Advisory Committee on Rules of Civil Procedure. The second critique was published by Professor David Freiman Engstrom in an issue of the Stanford Law Review dedicated to the 2012 Conference on Empirical Legal Studies.

Engstrom generously lauds the behavioral framework introduced in Locking the Doors. And Cecil professes “awe” at my purported “skill in extending a common theoretical economic model to the pretrial litigation setting.” But Engstrom suggests that the empirical work in Locking the Doors “is less sure-footed” than the paper’s conceptual framework, and Cecil launches a broadside not just against my use of the data in the two FJC reports he co-authored, but also against the very idea of using a model-guided framework for understanding the effects of Twombly and Iqbal.

To engage all of these critiques in detail here would render this Article un-
However, I shall respond to Cecil’s critiques of various aspects of the modeling approach itself, because this response fits with my focus on the broader methodological importance of taking motivated behavior seriously. As I discuss, Cecil’s attack on the idea of using a model of litigant behavior to guide empirical work is marred by two apparent misunderstandings. First, contra Cecil, the model is not a knock-off of the Priest-Klein framework, as Cecil claims. But more importantly, I shall show that Cecil mistakenly equates expositional assumptions that are made to simplify the discussion of the model with substantive limitations on the framework’s empirical implementation. I shall also respond to one aspect of Engstrom’s critique, concerning the proper unit of analysis in studying the empirical effects of Twombly and Iqbal. Before I address these critiques, however, I shall nail down some key methodological points.

II. Behaviorally Indifferent Empirics Has urned its Retirement

Generations of statistics students have been introduced to ideas related to statistical sampling using one form or another of an “urn problem.” In a simple example of such a problem, a collection of B blue balls and R red balls are contained in an urn; see Figure 1 for a picture of an urn. For some reason a statistician is interested in determining the fraction—call it \( \pi \)—of balls that are blue. A ball is selected—sampled, to be more clinical—from the urn, its color is recorded, it is put back into the urn, and this procedure is repeated \( N \) times. It is usually assumed that sampling of balls from the urn is done (i) randomly, so that every ball has the same probability of selection on every draw, and (ii) independently, so that what happens on one draw provides no information about what will happen on the next draw, given the numbers of red and blue balls in the urn. The recorded data are used to estimate \( \pi \), usually by taking the number of sampled balls that are blue, and dividing it by the total number of draws from the urn. This estimator, known as the sample proportion, has desirable properties: it is unbiased and consistent for the true proportion \( \pi \), as

---

65. Interested readers should see Jonah B. Gelbach, Can We Learn Anything About Pleading Changes From Existing Data? INTL. REV. L. & ECON. (forthcoming), where I respond to these arguments in detail.

66. See Cecil, Waves, supra note 13, at 38 & n.137, 44 (citing George L. Priest & Benjamin Klein, The Selection of Disputes for Litigation, 13 J. LEGAL STUD. 1 (1984)) (stating that Locking the Doors “is derived” from the “Priest/Klein model of litigation” and that Locking the Doors “adapts a common economic model of litigation”).


69. That is, \( \pi = B/(B+R) \).

70. An estimator \( \hat{\pi} \) is unbiased for population value \( \pi \) if \( E(\hat{\pi}) = \pi \), i.e., if the estimator’s expected value equals the population value. This means that on average, the estimator equals the population value.
Roughly speaking, an estimator $\hat{\pi}$ is consistent for population value $\pi$ if it is asymptotically unbiased, which means that as the sample size grows without bound, the probability converges to 0 that $\hat{\pi}$ and $\pi$ will differ by more than a stated magnitude, for any choice of that magnitude (formally, $P(|\hat{\pi} - \pi| > \epsilon)$). Consistency is a frequently used property because, while many estimators are biased in finite samples, it can be shown that their bias converges to zero as the sample size grows. An implication is that when the sample size is large enough, bias will be trivially small. How large is large enough is a tricky question, and one that is beyond the scope of this paper. The sample proportion is a special case in which consistency follows because of unbiasedness. More generally, consistency of estimators follows, when it does, from artful application of one or another law of large numbers. See, e.g., HALBERSTADT WHITE, ASYMPOTIC THEORY FOR ECONOMETRICIANS (2001).

An estimator $\hat{\pi}$ is asymptotically normal if, when its mean is $\pi$, the probability distribution of $\sqrt{N}(\pi - \pi)$ converges to a normal (bell curve) distribution as the sample size $N$ increases without bound. Roughly speaking, this means that for large enough sample sizes, one can use the normal distribution to analyze the behavior of the statistic $X$ even when $\hat{\pi}$ itself is not normally distributed for any given sample size $N$. Asymptotic normality of an estimator usually follows, when it does, from artful application of one or another central limit theorem. See, e.g., id.

An estimator is efficient in a class if it has the least variance among all estimators in that class. An estimator such as the statistic $X$ in note 72, supra, is asymptotically efficient in a class if, as the sample size increases without bound, its variance converges to a value that is the least such convergent value among all estimators in the class in question. Under simple random sampling, the asymptotic efficiency of the urn sample proportion follows from the fact that the sample proportion can be shown to equal the maximum likelihood estimator, which is well known to achieve the lowest possible asymptotic variance in the class of all consistent estimators. See, e.g., WILLIAM H. GREENE, ECONOMETRIC ANALYSIS 493 (2008).
One useful point of emphasis is that even in the simple urn-sampling case, we needed some assumptions about the selection mechanism used to choose balls to say much of anything statistically sensible. On the other hand, with the assumptions of random and independent sampling, we can say quite a lot about this proportion. A first lesson, then, is that statistical assumptions—by which I mean assumptions about the way observed data find their way into the analysts’ sights—can be very important.

When the commonly used baseline assumption of random sampling cannot be maintained, learning much at all about the object of estimation requires some knowledge concerning the actual way in which observations wind up in the researchers’ hands. For example, there are good reasons to believe that those researchers who create their Twombly/Iqbal study samples using searches of electronic databases may be sampling in nonrandom ways.

Violations of the independence assumption also can have important effects, though I shall not dwell on them here.

A second question, though, is: so what—who cares how many balls in an urn are blue or red? The urn problem almost invariably is a modeling construct. Just as law professors, practicing lawyers, and judges use fanciful hypotheticals to isolate legal issues of special interest, good statisticians use the urn problem, and others like it, to abstract from aspects of complex and contextualized real-world questions that are not fundamentally statistical in nature, in order to focus on the aspects of those questions that are amenable to statistical analysis. No one really cares about blue balls in urns. Rather, we represent real-world objects of interest, many of them abstract, using simple statistical analogues, because doing so clarifies the aspects of an empirical study that involve randomness, estimation, and inference. But human beings are not blue or red balls. They have motivations, and they have agency, and that has important implications for empirical research, including empirical research on legal pol-

74. To be precise, even without the random-sampling assumption, we can definitely say at least one thing, and possibly two things. First, if a blue ball is drawn at least one time, then we can say for sure that the true fraction of blue balls in the urn is at least as great as 1/(1+R). Second, if at least one red ball is drawn, then we can say for sure that the true fraction of blue balls is no greater than B/(B+1). These values operate, respectively, as lower and upper bounds on the true proportion π. The example as I laid it out involved sampling with replacement (we put the selected ball back in the urn after recording the result of each draw). If sampling had been done without replacement, then lower and upper bounds would be more informative, since the minimum number of blue balls in the urn would have to be at least the number of selected blue balls, and analogously for the number of red balls.

75. See FJC Initial Report, supra note 33; Cecil, Waves, supra note 13; Moore, Updated Quantitative Study, supra note 13; Engstrom, Twombly Puzzle, supra note 13.

icy. Thus, failing to take human behavior’s motivated agency into account—
treating outcomes of human behavior as simply so many urn-sampled balls—
can be fatal to the relevance of empirical research concerning legal rules and institutions.

To be clear, I do not mean to criticize the classical urn approach to statistical analysis. The vast majority of quantitative empirical work will have some irreducibly statistical component. Part of the art of doing good empirical work is figuring out where the behavioral model stops and the purely statistical work takes over. But a failure to take seriously the behavioral nature of data on humans can be disastrous in empirical research. To illustrate this point, I now turn to a classic example that every student of econometrics confronts.

A. Behavioral Challenges Illustrated: Estimating a Demand Function

Suppose the price of some good is given by $P$ and the logarithm of quantity demanded is given by $Q_D$, and suppose the good’s demand function is

$$ Q_D = \alpha + P \times \beta + \varepsilon_D. $$

Here, $\alpha$ and $\beta$ are, respectively, the demand function’s intercept and slope parameters, and $\varepsilon_D$ is a random element that captures variation in quantity demanded arising from unobserved factors. The theory of demand tells us that, except in freak cases, the parameter $\beta$ can be expected to be negative—when a good’s price rises, people generally want to buy less. Let us further assume things like the estimation of the population mean of a continuous random variable—something that is relevant when estimating the average level of income in a country, or the average level of damages awarded to prevailing tort plaintiffs in litigation. See, e.g., David A. Hyman, Bernard Black, Charles Silver, & William M. Sage, Estimating the Effect of Damages Caps in Medical Malpractice Cases: Evidence from Texas, 1 J. Legal Analysis 355 (2009). And many questions that involve discrete random variables, such as the binary outcome of whether an appellate judge votes for the plaintiff or defendant, usually are modeled as involving numerous covariates—not just the simple up-or-down outcome that could be directly analogized to the urn problem. See e.g., Christina L. Boyd, Lee Epstein, & Andrew D. Martin, Untangling the Causal Effects of Sex on Judging, 54 Am. J. Poli. Sci. 389 (2010); Joshua B. Fischman, Interpreting Circuit Court Voting Patterns: A Social Interactions Framework, J.L. Econ. & Org. (forthcoming).

77. I suspect that every empirical study can be viewed as dividing its analysis into separable components that involve purely statistical and purely substantive, non-statistical aspects even though users of statistics of course study more than urns. They also study things like the estimation of the population mean of a continuous random variable—something that is relevant when estimating the average level of income in a country, or the average level of damages awarded to prevailing tort plaintiffs in litigation. See, e.g., David A. Hyman, Bernard Black, Charles Silver, & William M. Sage, Estimating the Effect of Damages Caps in Medical Malpractice Cases: Evidence from Texas, 1 J. Legal Analysis 355 (2009). And many questions that involve discrete random variables, such as the binary outcome of whether an appellate judge votes for the plaintiff or defendant, usually are modeled as involving numerous covariates—not just the simple up-or-down outcome that could be directly analogized to the urn problem. See e.g., Christina L. Boyd, Lee Epstein, & Andrew D. Martin, Untangling the Causal Effects of Sex on Judging, 54 Am. J. Poli. Sci. 389 (2010); Joshua B. Fischman, Interpreting Circuit Court Voting Patterns: A Social Interactions Framework, J.L. Econ. & Org. (forthcoming).

78. Economists sometimes use the logarithm of quantity for a variety of reasons. Here I do so since it allows the measure of quantity to take on any real-number value, allowing the random element to have a normal distribution; this assumption is purely for exposition—dropping it would complicate the discussion without changing any substantive conclusion.

79. The freak cases are those in which the good is a Giffen good. See Alfred Marshall, 3 Principles of Economics Ch.VI, ¶ III.VI.17 (1895). In contemporary terms, imagine a poor law student with $21 to spend per week on dinner. He eats ramen noodles 6 nights a week, at a price of $2 per box, and on the seventh night he eats a $9 pizza. Now imagine a labor strike in the ramen industry causes a supply constriction, raising the price of ramen
that all data are observed at the market level—say, by counties—and that the random element $\epsilon$ is normally and independently distributed across counties, with mean zero. Finally, let us assume that we have a random sample of county-level data on the quantity and price of our good.

Along comes a statistician well-trained in the urn approach—the approach of modeling observed data in terms that (i) fit textbook statistical models but (ii) do not account for the human, behavioral factors that help determine the values of observed variables. Such a statistician might use the ordinary least squares estimator—sometimes called “running a regression”—to estimate $\alpha$ and $\beta$, noting on the side that she is making the usual assumption that the random element $\epsilon_D$ is uncorrelated with the independent variable (here, price). For simplicity, suppose the statistician observes data for just two types of counties. In counties of type 1, observed quantity and price are given by $Q_1$ and $P_1$, as in Figure 2. In counties of type 2, observed quantity and price are given by $Q_2$ and $P_2$. The result of our statistician’s use of ordinary least squares would be to determine that the estimated relationship between quantity and price is given by the dashed line in Figure 2. Assuming the estimated slope is statistically significant, our statistician would probably conclude that the data reject the theory of demand, since the estimated line has a positive slope, which is the wrong sign under the theory. And our statistician would be committing a major mistake.

noodles to $3 per box. After purchasing 6 boxes of ramen noodles at this price, our law student now has only $3 left for the seventh night—not enough for a pizza, but just enough for another box of ramen. Thus, an increase in the price of ramen noodles from $2 to $3 causes the student’s quantity demanded to rise from 6 boxes to 7. (Notice that this example depends importantly on the absence of some alternative to ramen—if mac ’n cheese were $2 per box and unaffected by the ramen strike, perhaps our student would shift demand from ramen to mac ’n cheese when the former’s price rose.)
The problem is that demand is just half the story. Economists assume that supply matters, too, and that prices and quantities are determined jointly by supply and demand. Suppose, again for exposition only, that the supply function is also log-linear:

$$Q_S = \gamma + P \times \delta + \epsilon_S,$$

where $Q_S$ is the logarithm of the quantity supplied, $\gamma$ and $\delta$ are, respectively, the supply function’s intercept and slope parameters, and $\epsilon_S$ is a random element that captures variation in quantity demanded arising from unobserved causes.

If market forces are allowed to determine price and quantity, then the price will be determined where the quantities demanded and supplied are equal. After setting quantity demanded and quantity supplied equal, it can be shown that among other things, the equilibrium price depends on the random element $\epsilon_D$.\(^80\)

\(^{80}\) Specifically, it can be shown that whenever $Q_D = Q_S$, we will have
In particular, when quantity demanded is higher for random reasons—when $\epsilon_D$ is greater—the demand curve shifts out, pushing up both equilibrium price and equilibrium quantity, as illustrated in Panel (a) of Figure 3. In this graph, the demand curve is negatively sloped, but there is a positive relationship between observed price and quantity because county 2’s demand curve has shifted out more than its supply curve, by comparison to county 1.

**Figure 3: Two Hypothetical Cross-Market Relationships Between Quantity and Price, in Which the Relationships Between Price and Quantity Supplied/Demanded Are the Same**

Now consider Panel (b) of Figure 3. There, we see the opposite situation from Panel (a): county 2’s supply curve has shifted out more than the demand curve has. Consequently, in Panel (b) we observe a negative relationship between price and quantity. But notice that all demand curves in the two Panels have the same slope, as do all supply curves. Thus, neither Panel’s observed relationship between price and quantity tells us anything useful about the parameters of the demand and supply curves. It simply makes no sense to use this sort of approach to test the theory of demand, or to estimate demand and supply curves.

There are really two aspects of the problem here. First, our statistician’s objective is poorly defined. It is unclear what it means to study “the causal re-

$$P = \frac{\alpha - \gamma}{\delta - \beta} + \frac{E_D - E_S}{\delta - \beta},$$

so that greater values of the demand shock $E_D$ are associated with greater values of equilibrium price, while greater values of the supply shock $E_S$ are associated with lower equilibrium price values.

81. That is, about $\alpha$, $\beta$, $\gamma$, or $\delta$. 
relationship between price and quantity” in the absence of a behavioral framework within the theory of supply and demand—or, perhaps, some other behavioral model that the researcher is willing to defend in explicit terms.

The second aspect of the problem is evident once we posit a clear behavioral framework—in this case, the theory of demand and supply in competitive markets. The discussion above—and especially Figure 3—show that the simple relationship between observed county-level quantity on observed county-level price cannot identify either the demand curve or the supply curve. Only by finding a way to deal with unobserved heterogeneity in supply and demand—the $\varepsilon_D$ and $\varepsilon_S$ terms above—can an empirical researcher hope to validly estimate the demand and supply curves in a competitive market. Since these curves work together to determine price and quantity, only an approach that accounts for both supply and demand can uncover the actual causal relationships that connect the observed variables of price and quantity.

To be clear, the problem here isn’t that demand and supply are never estimable; economists long ago devised methods to deal with the unobserved heterogeneity problem, and these methods have been among the core techniques taught in econometrics courses for many decades.\(^\text{82}\) Instead, the problem is a mismatch between the urn approach to statistical analysis and the human, behavioral choices that determine the values of the observed variables. The takeaway point here is that properly understanding the empirical determinants of price and quantity requires more than just a theoretical statistical analysis—it requires serious attention to the behavioral origins of data on humans.

Similarly, empirical facts about the civil justice system may fail the relevance test if they are not interpreted within a behavioral framework. Thus, a critical step in policy-relevant empirical work is to consider clearly what behavioral possibilities are analytically admissible. In the foregoing discussion, for example, the behavioral possibilities include the possibility (to most economists, the inevitability) that a good’s price and quantity will be determined together by the interaction of producers and consumers as reflected by supply and demand. Similarly, in the case of studying the effects of Twombly and Iqbal, as I shall discuss in the next Part, the general set of possibilities includes the specific possibility that parties will change their behavior in response to perceived changes in the pleading standard.

When a researcher allows for a particular behavioral possibility, she signs up for the obligation to develop an identification strategy that is capable of un-

\(^{82}\) The usual approach is to find observable instrumental variables, each of which plausibly shifts either the demand curve or the supply curve, but not both. An instrument that shifts the demand curve helps identify the parameters of the supply curve, while one that shifts the supply curve helps identify the parameters of the demand curve. See, e.g., Joshua D. Angrist & Alan B. Krueger, Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments, 15 J. ECON. PERSP. 69 (2001), for an accessible discussion of instrumental variables in this context.
covering useful empirical information when the behavioral possibility in ques-
tion might occur. The term “identification strategy” is widely used in applied
econometrics, but so far as I know it does not have a widely agreed formal
definition. To be clear, by “identification strategy” I shall mean a set of behav-
ioral assumptions, statistical assumptions, data sources, and estimation meth-
ods that can be used together to estimate an empirical object of interest.

It is important to recognize that stronger behavioral assumptions may allow
researchers to claim identification of more objects of empirical interest. Con-
versely, it is often the case that a researcher’s claims of identification rely on
strong behavioral assumptions. This latter point is especially important to rec-
ognize because the key behavioral assumptions are all too often left unstated.
Consider again our hypothetical statistician who wants to estimate the causal
relationship between price and quantity. As we saw above, when the theory of
demand and supply is correct, the observed relationship between price and
quantity generally will not identify either the demand curve or the supply
curve. Once again, the problem is that observed co-movements in quantity and
price are affected by unobservable factors causing market-level variation in the
location of both curves: variation in the intensity of potential buyers’ desires for
the good shift the demand curve, while variation in production costs shift the
supply curve.

There are, however, special cases under which the statistician’s estimation
approach described above does yield valid estimates of something of interest.
For example, suppose there is no unobserved heterogeneity across counties in
the location of the demand curve—which is to say, no variation in \( \varepsilon_D \). In that
case, the only source of variation in price and quantity must be cross-county
shifts in the supply curve. As Figure 4 illustrates, the observed relationship be-
tween price and quantity identifies the demand curve in this situation. Thus, a
statistician who claims to test the theory of demand by using only data on the
observed cross-market price-quantity relationship could justify this approach by
clearly embracing the assumption that the location of the demand curve does
not vary across counties. In general, few would consider such an assumption
reasonable. And that is one reason why it is so important to be clear about beha-
vioral assumptions: strong and even unreasonable assumptions are often hid-
en, lurking just beneath the veneer of an intuitive-seeming empirical approach.

83. To be even clearer, by “statistical assumptions” I mean assumptions that character-
ize “all of the random influences that combine together to lead to individual observations”
(this specific quotation comes from Data Generating Process, WIKIPEDIA,
Consider William Hubbard’s two studies on Twombly and Iqbal. Hubbard claims to estimate the effect of Twombly and Iqbal on dismissal rates—what I call judicial behavior effects—and he claims to do so while “controll[ing] for selection effects.” In his first study, Hubbard attempts to study the effects of Twombly alone (i.e., he does not try to study Iqbal’s effects) by comparing (i) measures of the grant rate using only cases filed in the forty-five-day period just before the Supreme Court released the Twombly decision to (ii) measures of the outcomes for cases filed in the same calendar period a year earlier. Hubbard’s reason for isolating these “straddle cases” is that “if the plausibility standard announced in Twombly led many plaintiffs not to file suit at all, it is possible that the share of filed cases being dismissed may not change, even though many (potential) plaintiffs are nonetheless losing their day in court.”

84. See Hubbard, Testing for Procedural Change, supra note 40; Hubbard, Theory of Pleading, supra note 40, at 16.
86. So far as I know, the term “straddle cases” was coined by Engstrom, supra note 61, at 1224.
But Hubbard’s approach does nothing to eliminate the problems of defendant selection and settlement selection, since Twombly’s occurrence was knowable to all parties within a short period of time after case filing. As I show in Part III.B, infra, these types of selection are themselves sufficient to render the change in the grant rate generally uninformative about changes in judicial behavior. In other words, eliminating plaintiff selection is an insufficient identification strategy, on its own, to tell us anything of interest about Twombly and Iqbal’s effects.

As appealing as Hubbard’s approach might seem, it controls for only plaintiff selection effects. Thus Hubbard’s claim to having identified judicial behavior effects relies on the important, unstated assumption that defendant selection and settlement selection effects either don’t exist or don’t matter empirically if they do. This example illustrates the dangers of failing to clearly consider and state the assumptions guiding empirical work generally, and thus also in civil procedure: the fact that assumptions aren’t discussed doesn’t mean they aren’t doing important work.

B. The Absence of Party Selection as a Behavioral Restriction

The problem of hidden assumptions in estimating the effects of Twombly and Iqbal is certainly not limited to the point I have just made about Hubbard’s work. Any identification strategy directed at learning something meaningful about pleading standard changes using only the change in the grant rate necessarily includes behavioral restrictions on the extent of party selection. Consequently, researchers who claim that changes in the grant rate, on their own, tell us something useful about the effects of Twombly and Iqbal must shoulder the burden of explaining the behavioral restrictions on party selection effects on which they are relying. Unfortunately, this has not generally been the way of things in the literature on Twombly and Iqbal. I have already noted, for example, that William Hubbard’s approach plausibly eliminates only plaintiff selection effects, leaving implicit his apparent assumption that there are neither defendant selection nor settlement selection effects.

Cecil takes an even more problematic approach. He states that he has “no quarrel” with the proposition that there is defendant selection, and he also al-

88. That said, it plaintiffs could voluntarily dismiss their suits via Rule 41 when faced with Rule 12(b)(6) motions that they did not expect to face at the time they filed suit; thus it is possible that some plaintiff selection effects remain even in Hubbard’s data.

89. I provide a detailed discussion of “surprise” cases, in which the parties are surprised to find a new pleading standard in place after the plaintiff files suit, in Part V of Jonah B. Gelbach, Can We Learn Anything About Pleading Changes from Existing Data? INTL. REV. L. & ECON. (forthcoming).

90. See Part A, supra.

91. Cecil, Waves, supra note 13, at 42.
allows the possibility that there are both plaintiff and settlement selection effects that might occur, and he does not make any assumptions that would restrict the magnitude of any of them. But Cecil also insists that in using the change in the grant rate to measure the effects of Twombly and Iqbal, his co-authored studies “provide[ ] the best estimate[s] of the federal district courts’ responses to these decisions.” Here one can only cry, foul. Given that one has accepted the empirical correctness of a premise, logic dictates that one must also accept as sound the conclusion of any valid argument starting from that premise. Something must go: either Cecil must withdraw his endorsement of grant rate comparisons, or he must endorse a set of restrictions on party selection strong enough to make such comparisons informative.

For readers who are unfamiliar with thinking through the role of behavioral assumptions in identifying empirical objects of interest, perhaps it will be helpful to be a bit more concrete. Here is a simple statement of one possible behavioral restriction: parties do not change their behavior following perceived changes in the pleading standard. That is, the pleading standard has no impact on whether plaintiffs file suit, nor on whether defendants file Rule 12(b)(6) motions in cases plaintiffs have filed, nor on whether cases settle before the Rule 12(b)(6) stage. As I discuss in Part III.B, if this restriction is true, the change in the grant rate could tell us something important—namely, how judicial behavior has changed as a result of Twombly and Iqbal.

I do not think anyone seriously believes this “no-selection” behavioral restriction to be accurate; certainly, no one has proudly pledged allegiance to it.

To be clear, I do not mean to suggest that behavioral restrictions are unacceptable—quite the contrary. Such restrictions are unavoidable if we are to use data to learn about the world: if absolutely any behavioral response to a change in the pleading standard could happen, then there will be no way to learn anything about its effects from any data. Behavioral assumptions are necessary, and we all make them all the time in our daily lives. For example, you cross a heavily trafficked street when the walk sign is illuminated because you think that motorists will not run a red light. That is a behavioral assumption, and an obviously important one—if you’re wrong, very bad things very possibly will happen to you.

One strand of the discussion here may be summarized using what I call the “no blood from a behavioral stone” principle of empirical work. Without any assumptions on human behavior, one simply cannot learn anything of policy

---

92. Cecil, Waves, supra note 13, at 42 & nn. 138-39 (“[A]n increase in motions by defendants may be caused by a drop in the rate at which cases settle. . . . [O]ur ‘filings’ dataset would also include any plaintiff selection effects. . . . Sorting out the contributions of the ‘plaintiff selection effect’ and the ‘settlement selection effect’ is a difficult task.”).


94. To be sure, other assumptions would still be necessary, such as that the case mix did not change, in empirically important ways, for reasons unrelated to changes in the pleading standard.
importance. Of course researchers should strive to make the least restrictive assumptions that allow empirical research to answer normatively relevant questions. But statements like “that method requires assumptions” are beside the point, because any empirically informative approach requires assumptions. The alternative to one set of assumptions is not no assumptions, but rather some other assumptions. We need more clarity from researchers about the behavioral restrictions and assumptions embedded in their mapping of empirical facts into normatively relevant ones, because the appropriateness of often untestable behavioral assumptions can be productively debated only when these assumptions are allowed out in the open.

The second strand of this discussion is a bit more subtle, and is actually prior to the no blood from a stone principle. One might call it the “understand which question you’re studying” principle. There’s little point in arguing over whether an estimate is large or small, or larger or smaller than some other estimate, if no one has a clear understanding of the estimand—i.e., the object of estimation. It seems likely that most authors of empirical studies of Twombly/Iqbal have had in mind estimating something like what Hubbard has called the “true effect on dismissal rates.”

Using this phrase makes it seem like all researchers are trying to estimate some particular parameter’s value, and that any problems in so doing are the result of some sort of estimation challenge, whether due to party selection effects or other factors such as changes in the economy that change litigation behavior. Yet that can’t be right, for at least two reasons. First, there is surely variation across cases in the probability of dismissal: some complaints are stronger than others; some cases are assigned to judges who may be more defendant- or plaintiff-friendly. Moreover, as a practical matter some cases will never face Rule 12(b)(6) motions, since plaintiffs in those cases will have no problem pleading plausibly. Since we never observe such cases facing Rule 12(b)(6) motions, we have no way (other than via assumption) to learn anything at all about “the true” dismissal rate for cases such as these. In the applied econometrics literature, much attention has been paid in recent decades to what has come to be known as treatment effect heterogeneity. The issues can get very technical and notationally dense, but roughly speaking, the end result

---

95. Hubbard, A Theory of Pleading, supra note 40, at 15.
96. James Heckman has made the point that certain types of overall effects are neither policy relevant nor interesting to observe. See, e.g., James J. Heckman, Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations, 32 J. HUM. RESOURCES 441, 443-44 (1997) (“Picking a millionaire at random to participate in a training program for low skilled workers, or making an idiot into a PhD may be intriguing thought experiments but are usually neither policy relevant nor feasible. They are not policy relevant because interest centers on the effects of programs on intended recipients—not on persons for whom the program was never intended. It is not a feasible random-assignment strategy because millionaires would never agree to participate in such a training program even if they were offered the chance to do so, and few idiots would be able to attain the PhD in most fields.”).
has been an agreement that what researchers typically are estimating in such circumstances—their estimand—is some sort of average effect among units that could be affected by the policy change at issue.\footnote{97}{See, e.g., James Heckman & Richard Robb, Alternative Methods For Evaluating the Impact of Interventions, in LONGITUDINAL ANALYSIS OF LABOR MARKET DATA 156-245 (James Heckman & Burton Singer eds., 1985); Guido W. Imbens & Joshua D. Angrist, Identification and Estimation of Local Average Treatment Effects, 62 ECONOMETRICA 467 (1994). For studies seeking to identify distributional effects, rather than simply mean impacts, see, e.g., John DiNardo, Nicole M. Fortin, & Thomas Lemieux, Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach, 64 ECONOMETRICA 1001 (1996); James J. Heckman, Jeffrey Smith & Nancy Clements, Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts, 64 REV. ECON. STUDIES 487 (1997); Marianne P. Bitler, Jonah B. Gelbach, & Hilary W. Hoynes, What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments, 96 AM. ECON. REV. 988 (2006).}

Thus it is important to understand that even if Hubbard’s “true effect on dismissal rates” could be identified empirically, it would be some sort of average of the true effect on the probability of dismissal for some subset of all disputes. In Hubbard’s context, that subset would be something like “the set of disputes that would turn into lawsuits if the parties expected one pleading standard to govern when the plaintiff filed suit but then found out some other standard would govern.” At best, Hubbard’s approach identifies an average effect among such disputes, which might differ importantly from other disputes of interest. To be sure, the negatively affected share measure I discuss in Part III.C, infra, also concerns only a subset of cases (those that would have a Rule 12(b)(6) motion filed under the post-Iqbal pleading standard). Thus I do not mean to criticize Hubbard, or anyone else, for using a method that could identify policy impact for only some types of cases; without further assumptions that ensure homogeneity of policy effects, this is simply a fact of empirical life.

Second, and at least as important, is the fact that “the probability of dismissal” is itself a problematic and incomplete object of study. As I argued in Locking the Doors and shall discuss in Part III, infra, a change in the pleading standard can be expected to affect parties’ welfare through direct selection-related channels, not just through the change in the probability that any given case would be dismissed. Thus, it is a mistake to compare the results in Locking the Doors—which self-consciously considered a broader array of effects than simply changes in judicial behavior—directly to others in the literature.\footnote{98}{For example, Hubbard commits this error when he writes that Locking the Doors “account[s] for selection effects in estimating the effects of Twombly or Iqbal on dismissal rates.” Hubbard, Testing for Change, supra note 40, at 45.} This example shows that it is important not only to be clear about the behavioral framework in which an estimate is to be interpreted, but also to be clear about what object is being estimated, and why that object is of interest.
In this Part I discuss the behavioral framework I built and explored in *Locking the Doors*, as well as its connection to the empirical results reported there. A key building block in *Locking the Doors* is the idea that for any pleading standard, each dispute has a potential outcome that would obtain in that dispute if the pleading standard in question governed. It is critical to recognize that potential outcomes are counterfactual objects. Regardless of what does happen, potential outcomes tell us something about what would happen if a state of the world were to occur.

In discussing the behavioral framework in *Locking the Doors*, I assumed that each case involves a single plaintiff with a single claim against a single defendant, that Rule 12(b)(6) motions are the only type of motion to dismiss, and that any time such a motion is granted, it is granted without allowing the plaintiff leave to amend her complaint. For exposition’s sake, I shall at first retain these assumptions; in Part III.F.1, infra, I explain why their importance is only expositional—no qualitative differences arise when I relax these assumptions.

### A. Potential Outcomes

In *Locking the Doors*, I noted that we can separate cases into several categories of potential outcomes:

- “D” disputes: those that are dropped without the plaintiff’s filing a complaint;
- “S” disputes: those that involve an agreed settlement before the defendant files either an answer or a motion to dismiss (these cases might be settled either before or after the plaintiff files her complaint);
- “A” disputes: those in which the plaintiff files a complaint and the defendant files an answer without filing a motion to dismiss;
- “M” disputes: those in which the plaintiff files a complaint and the defendant files a motion to dismiss.

I then observed that since each dispute has a unique potential outcome under a given pleading standard, we can further categorize disputes according to the sixteen logically possible pairs of these potential outcomes that might obtain under the *Conley* and *Twombly/Iqbal* pleading standards. The matrix in Figure 5 illustrates these sixteen possible dispute types, with potential outcomes under *Conley* represented in the matrix’s rows and potential outcomes under *Twombly/Iqbal* represented in its columns.

---

Thus, for example, a dispute of type DD is one that would be dropped by the plaintiff under Conley (the first “D”) and also dropped under Twombly/Iqbal (the second “D”), while an AA dispute is one in which, under both pleading standards, the plaintiff would file suit and the defendant would file an answer; similarly, SS disputes are settled under both pleading standards and MM disputes involve a plaintiff’s filing suit and a defendant’s filing a motion to dismiss. These four types of disputes—whose potential outcome is the same under the two pleading standards, and which are represented with shaded cells on the top-left to bottom-right diagonal part of the Figure 5 matrix—are what I term “non-selection” cases. Disputes with different potential outcomes under the two pleading standards are those that involve some sort of selection, since parties do different things in these disputes as a function of the governing pleading standard. These dispute types are represented by the cells that lie off the diagonal that runs from the top-left to the bottom-right of Figure 5’s matrix. I define settlement selection to include any dispute that would be settled under one pleading standard but not the other. For example, in SM disputes, the parties would settle under Conley, but under Twombly/Iqbal the plaintiff would file suit and the defendant would file a Rule 12(b)(6) motion; thus SM disputes in-
volve settlement selection. All told, the dispute types that involve settlement selection are SD, SM, DS, AS, and MS. Settlement selection dispute types are represented in Figure 5 using bold font in the cell labels.

Dispute types that do not involve settlement selection but do involve a change in whether the defendant files a Rule 12(b)(6) motion are those that I term defendant selection disputes. For example, AM disputes involve defendant selection, because while the plaintiff would file suit under either standard, the defendant would file an answer under Conley but file a motion to dismiss under Twombly/Iqbal. In Figure 5, defendant selection disputes are represented with underlined cell labels. Finally, I say that disputes involve plaintiff selection if they do not involve settlement selection but do involve a change in whether the plaintiff files suit. Thus plaintiff selection occurs in disputes of type AD, MD, DA, and DM. Dispute types in which there is plaintiff selection are represented with italic font in the cell labels in Figure 5.

The next step in Locking the Doors was to divide “M” cases into two more refined potential outcome sub-categories:

- “M₀” cases: those M cases in which the defendant’s motion to dismiss would be denied;
- “M₁” cases: those M cases in which the defendant’s motion to dismiss would be granted.

Now we have five, rather than four, potential outcomes for each pleading standard. Since each dispute has one potential outcome under each pleading standard, there are 25 logically possible dispute types: each of the five types of potential outcome under Conley, paired with each of the five types under Twombly/Iqbal, as depicted in the expanded matrix in Figure 6. The matrix retains Figure 5’s approach of using different shading and fonts to highlight which dispute types are which.

---

100. Notice that dispute types DM and MD involve both plaintiff and defendant selection.
### Figure 6: Expanded Taxonomy of Cases in Terms of Potential Outcomes Under the Conley and Twombly/Iqbal Pleading Regimes

<table>
<thead>
<tr>
<th>Outcome under Conley Standard</th>
<th>MTD Filed</th>
<th>Twombly/Iqbal Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropped</td>
<td>Dropped</td>
<td>Dropped</td>
</tr>
<tr>
<td>Settled</td>
<td>Settled</td>
<td>Settled</td>
</tr>
<tr>
<td>Answered</td>
<td>Answered</td>
<td>Answered</td>
</tr>
<tr>
<td>Denied</td>
<td>Denied</td>
<td>Denied</td>
</tr>
<tr>
<td>Granted</td>
<td>Granted</td>
<td>Granted</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MTD Filed</th>
<th>Twombly/Iqbal Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropped</td>
<td>Dropped</td>
</tr>
<tr>
<td>Settled</td>
<td>Settled</td>
</tr>
<tr>
<td>Answered</td>
<td>Answered</td>
</tr>
</tbody>
</table>

#### Legend
- **Shaded** — Non-selection disputes
- **Bold** — Settlement selection disputes
- **Underlined** — Defendant selection disputes
- **Italicized** — Plaintiff selection disputes

One feature of Figure 6 is that it helps clarify which judicial behavior effects might possibly be identified empirically. Formally, I define a given dispute’s *judicial behavior effect* to be (i) the probability that a judge would grant a Rule 12(b)(6) motion in that dispute under *Twombly/Iqbal* if the dispute were to reach the Rule 12(b)(6) stage of litigation under that pleading standard, minus (ii) the corresponding probability of a Rule 12(b)(6) grant under *Conley*, again if the dispute were to reach the Rule 12(b)(6) stage under that pleading standard.
Thus, the judicial behavior effect is defined for every dispute, even if the dispute would never actually have a Rule 12(b)(6) motion adjudicated. For example, consider a dispute over who won a round of Rock, Paper, Scissors," with stakes equal to $1, between a single plaintiff and a single defendant. Assuming neither party cares about the principle involved, this dispute would obviously never lead to a federal lawsuit, much less a Rule 12(b)(6) motion. But the question of what would happen under each pleading standard, if such a dispute were litigated to the Rule 12(b)(6) stage, is logically prior to whether the event will happen. So it is proper to speak of the judicial behavior effect for this dispute—and, indeed, for any dispute. That said, observe that the only dispute types in Figure 6 that would see Rule 12(b)(6) motions filed under both pleading standards are, by definition, those involving disputes that are Type M under both pleading standard. We can define the judicial behavior effect among these dispute types as

\[ JBE_{MM} = \frac{M_B M_G}{MM} \]

This parameter tells us the share of MM cases in which a Rule 12(b)(6) motion is granted only because of the switch to the Twombly/Iqbal pleading standard—that is, it tells us the share of MM cases in which Twombly/Iqbal but-for cause a Rule 12(b)(6) motion grant. Thus, JBE_{MM} is the kind of parameter that authors in the empirical literature seem to have in mind. In the next section, I analyze the circumstances under which JBE_{MM} can be identified empirically and conclude that these circumstances are unlikely.

First, though, we should take note of another important point: even if JBE_{MM} could be identified, this judicial behavior effect parameter would tell us the judicial behavior effect only among MM disputes. It is entirely possible that the effect would be different among other dispute types. This is important to recognize because it is parties’ perceived changes in judicial behavior that will be the driving force in changing party behavior. And many cases might involve party selection, placing them outside the MM set. Thus, even if we could learn the judicial behavior effect among MM cases, we would be missing much of the landscape that is relevant to understanding how party behavior changes. This is an example in which it is important to keep in mind Fischman’s admonition to “think more carefully about how empirical findings generalize from a research setting to a policy-relevant context.”


103. See supra discussion of treatment effect heterogeneity at notes 96-96.

104. Fischman, supra note 5, at 154.
B. Using an Economic Model as Behavioral Framework Eliminates the Possibility of Some Dispute Types, But Not Enough to Make Grant Rate Differences Informativethe discussion in section A, supra, imposed no particular behavioral framework: it considered all logically possible dispute types in terms of their pair of potential outcomes under the two pleading standards. Here I follow the approach in Locking the Doors: I sketch a simple economic model of pretrial litigation whose key foundations are the assumptions (i) that parties behave rationally, in the sense that each party pursues her self-interest as she understands it, and roughly speaking, and (ii) that litigation activity occurs if and only if at least one party expects to be better off litigating than either settling or allowing the case to move to the next stage.\footnote{I work out the details of this model in considerable mathematical detail in Gelbach, Selection in Motion, supra note 17.}

One payoff from using this model is that we can eliminate from consideration the possibility that there are disputes of certain of the types discussed above. First, because there is no reason to think Twombly/Iqbal actually reduce defendants’ chances of winning on a motion to dismiss, we can rule out $M_D M_D$ disputes; in such disputes a judge would grant the Rule 12(b)(6) motion under Conley but deny it under Twombly/Iqbal. Second, we can rule out $M_D A$ and $M_A A$ disputes: a defendant who would find it worth filing a motion to dismiss when her adversary files suit under Conley will never find it worthwhile to back down and file an answer under Twombly/Iqbal. We can rule out the presence of $D S$, $D A$, $D M_D$, and $D M_A$ disputes for essentially the same reason: whatever the precise effect of Twombly and Iqbal, they hardly could improve plaintiffs’ assessment of the returns to litigating. In Figure 7, I repeat the taxonomy from Figure 6, except that I indicate the dispute types whose presence is ruled out under the economic model by blacking out the cells that represent them.
It can be shown, however, that all dispute types whose cells in Figure 7 are not blacked out might occur even when parties are rational in the way discussed above. An important consequence is that the change in the Rule 12(b)(6) grant rate is broadly uninformative about any effect of interest. Observe the following:

---

106. See Gelbach, Selection in Motion, supra note 17.
• The dispute types that have a Rule 12(b)(6) motion granted under Conley include non-selection disputes (M₃M₄), settlement selection disputes (M₄S), and plaintiff selection disputes (M₄D).

• The dispute types that have a Rule 12(b)(6) motion filed under Conley are those that have grants under Conley, as well as M₃M₄, M₅M₆, M₄S, and M₄D disputes.¹⁰⁸

• Thus, using the convention that the number of disputes of each type is represented by the dispute type’s label, the grant rate under Conley is:

\[ g_C = \frac{M₃M₄ + M₄S + M₄D}{M₃M₄ + M₅M₆ + M₄S + M₄D} \]

• The dispute types that have a Rule 12(b)(6) motion granted under Twombly/Iqbal include non-selection disputes (M₃M₄ and M₅M₆), defendant selection disputes (AM₅), and settlement selection disputes (SM₅).

• The dispute types that have a Rule 12(b)(6) motion filed under Twombly/Iqbal are those that have grants under Twombly/Iqbal, as well as M₃M₄, AM₅, and SM₅ disputes.¹¹⁰

• Again using the convention that the number of disputes of each type is represented by the dispute type’s label, the grant rate under Twombly/Iqbal is:

\[ g_T = \frac{M₃M₄ + AM₅ + SM₅}{M₃M₄ + AM₅ + SM₅} \]

Obviously all this is a bit of a mess. And the difference between the grant rates \( g_C \) and \( g_T \) is even worse; it certainly does not seem that anything simple can be learned by measuring the difference in the grant rate across pleading standards. With a bunch of boring algebra, it can be shown that the difference in the grant rates equals:

\[ \epsilon \approx \frac{M₃M₄ + M₅M₆ + M₄S + M₄D}{M₃M₄ + M₅M₆ + M₄S + M₄D} \]

107. Thus the disputes that have a Rule 12(b)(6) motion granted under Conley include non-selection disputes (M₃M₄), settlement selection disputes (M₄S), and plaintiff selection disputes (M₄D).

108. Thus, in addition to the dispute types with Rule 12(b)(6) motions granted under Conley, the disputes that have a Rule 12(b)(6) motion filed under Conley—whether the motion is granted or denied—include additional non-selection disputes (M₃M₄ and M₅M₆), additional settlement selection disputes (M₃S), and additional plaintiff selection disputes (M₄D).

109. Thus, the disputes that have a Rule 12(b)(6) motion granted under Twombly/Iqbal include non-selection disputes (M₃M₄ and M₅M₆), defendant selection disputes (AM₅), and settlement selection disputes (SM₅).

110. Thus, in addition to the dispute types with Rule 12(b)(6) motions granted under Twombly/Iqbal, the disputes that have a Rule 12(b)(6) motion filed under Twombly/Iqbal—whether the motion is granted or denied—include additional non-selection disputes (M₃M₄), additional settlement selection disputes (SM₅), and additional defendant selection disputes (AM₅).
where $B_{1T}$, $B_c$, and $B_A$ all depend on the numbers of disputes of various types.\footnote{111} It can be shown that both $B_{1T}$ and $B_c$ are positive and no greater than 1, and that either can be the greater of the two. It can also be shown that $B_A$ can be either positive or negative.

I shall now illustrate the pathologies that inhere in using the difference in grant rates to measure the judicial behavior effect parameter $JBE_{MM}$. First suppose that Twombly and Iqbal have no impact on judicial behavior at all, so that $JBE_{MM}$ is zero. In that case, the change in the grant rate equals the term in square brackets just above. It is easy to construct examples in which this term is positive, negative, or zero.\footnote{112} And even if the judicial behavior effect is positive—so that judges dismiss more MM cases under Twombly/Iqbal than they would under Conley—it is possible to construct examples in which the difference of the grant rate takes on any sign.\footnote{113} Consequently, observed values of the change in the grant rate carry no information even about the sign of judicial behavior effects. In other words: by itself, the change in the grant rate tells us nothing at all about the judicial behavior effect among MM cases.\footnote{114} The

\begin{align*}
B_v &= \frac{MM}{MM + AM + SM}, \\
B_c &= \frac{MM}{MM + MS + MD}, \\
B_A &= \frac{AM + SM}{MM + AM + SM} \cdot \frac{MS + M_D}{MM + MS + MD}.
\end{align*}

The $B_v$ term tells us the share of cases with Rule 12(b)(6) motions filed under Twombly/Iqbal in which there is nor party selection, and the $B_c$ term tells us the corresponding fact for cases with Rule 12(b)(6) motions filed under Conley. The first part of the $B_A$ term measures the importance of party selection among cases with Rule 12(b)(6) motions granted under Twombly/Iqbal, as a share of the number of cases with such motions filed under Twombly/Iqbal; the second part of the $B_A$ term measures the analogous fact under Conley. Thus, $B_A$ is positive when selection is a proportionately more important factor in driving grants under Twombly/Iqbal than it is under Conley, with importance calculated proportionately to the number of motions filed under each respective pleading standard.

\footnote{112} Recall that ($B_v - B_{1T}$) may have any sign, as can $B_c$; it is also possible to show that neither the sign of ($B_v - B_{1T}$) nor the sign of $B_c$ restricts the other. By setting both ($B_v - B_{1T}$)\(>0\) and $B_c\ge0$, we get a positive square-bracket term (since the grant rate under Conley is always positive). By reversing the inequalities, we get a negative square-bracket term, and by replacing them with “=” we get a zero square-bracket term.

\footnote{113} See Table 1 of Gelbach, Locking the Doors, supra note 17 at 2313 (providing an example of each type, with a true judicial behavior effect among MM cases of 20 percentage points).

\footnote{114} In fact, I go further than this in other draft work. See Gelbach, Selection in Motion, supra note 17 (showing that both the presence and absence of judicial behavior effects is consistent with any combination of (i) change in the grant rate, (ii) change in the share of filed cases that face Rule 12(b)(6) motions, and (iii) number of cases filed).
A change in the grant rate is just like our hypothetical statistician’s attempt to test a theory of demand by regressing quantity on price (see Part II.A, supra), where any empirical finding would have been consistent with the empirical correctness of the theory.

A quick look at equation (1) shows that the difference in the grant rate does identify JBE\textsubscript{MM} when B\textsubscript{TI}=1, B\textsubscript{C}=1, and B\textsubscript{Δ}=0 all hold. What are the conditions under which these three conditions all are satisfied? Precisely the conditions under which there is no party selection. This can be seen by inspecting note 111, supra, and observing that

- B\textsubscript{TI}=1 if and only if there are no AM or SM cases (so that there is no selection into Rule 12(b)(6) motion filing);
- B\textsubscript{C}=1 if and only if there are no MS or MD cases (so that there is no selection out of Rule 12(b)(6) motion filing);
- B\textsubscript{Δ}=0 if each of the first two conditions holds.\textsuperscript{115}

Thus, just as our statistician from Part II can claim her regression results estimate the demand curve if she is willing to assume away variation in demand, so, too, can a civil procedure researcher claim to identify the judicial behavior effect among MM cases if she is willing to assert the absence of defendant selection (which eliminates AM disputes), plaintiff selection (which eliminates MD disputes), and settlement selection (which eliminates SM and MS disputes); notice that when there are no party selection effects of these types, there is no party selection at all.

I have made no bones about my view that it is entirely unreasonable to make this assumption. In part, that is because the empirical evidence decisively rejects the assumption that there is no party selection. For example, the FJC authors’ initial report shows that the rate at which Rule 12(b)(6) motions are filed rose substantially among filed cases.\textsuperscript{116} All else equal, such a result can be observed only if there are some disputes whose pair of potential outcomes has an “M” at the end but not at the front. In other words, it can be observed only if there is either defendant selection—AM cases—or settlement selection—SM cases.\textsuperscript{117}

This discussion shows that the assumption of no selection is rejected by the data. But that assumption was only a sufficient condition for the difference in the grant rate to identify the judicial behavior effect among MM cases. It is natural to wonder whether there are other assumptions that would yield identification of the judicial behavior effect parameter JBE\textsubscript{MM}. Formally, the answer is yes, but for practical purposes it is no, because the required assumptions both

\textsuperscript{115} Notice that B\textsubscript{Δ} can equal 0 even when B\textsubscript{TI} and B\textsubscript{C} are not both 1. However, I do not pursue this point since B\textsubscript{TI}=1 and B\textsubscript{C}=1 both require the absence of selection.

\textsuperscript{116} See FJC Initial Report, supra note 33, at tbl.1.

\textsuperscript{117} Under the restrictions of the economic model, there will be no DM cases.
are very strong and appear impossible to motivate in an intuitive way. Consequently, it plausibly eliminates plaintiff selection effects among MM cases are out of luck: the assumptions necessary to do so are too strong to accept.

This admonition applies to all empirical studies that have concentrated on using Twombly and Iqbal’s effects on Rule 12(b)(6) outcomes to measure judicial behavior effects. That includes Hubbard’s work, even though, as discussed supra, it plausibly eliminates plaintiff selection effects. In the framework discussed here, eliminating plaintiff selection effects means eliminating M_D and M_S cases. But this restriction has no effect on our B_IR term, and while it does limit the types of cases that might be represented in our B_C and B_A terms, it does not do so in a way that provides any more useful information about the relative magnitude of these terms.

118. After a lot of tedious algebra, it can be shown that when B_IR and B_C are not both 1, the difference in the grant rate equals the judicial behavior effect among MM cases if and only if

\[ JBE_{MM} = (1 - B_{IR}) \frac{(MS + MD) - (AM + SM)}{(MM + AM + SM)(MM + MS + MD)} M_D M_S + B_C. \]

The most important thing to note is that this high-level condition places a restriction on JBE_{MM}, the object of measurement itself: only when the judicial behavior effect happens to equal a complicated function of various selection-related effects will the grant rate identify JBE_{MM}. It might seem that some headway could be made via the assumption that, which implies the same number of cases select out of, and into, the state of having Rule 12(b)(6) motions as a result of Twombly and Iqbal. Under this restriction, the first term in square brackets would be zero. Even then, though, the difference in the grant rate identifies JBE_{MM} only if JBE_{MM} itself just happens to equal B_C/(1-B_{IR}), and it can be shown that this implies the restriction that JBE_{MM} equals the ratio [(AM+SM)+(M_S+M_D)]/(AM+SM), which isn’t even possible when the numerator term is negative; more generally, there is just no reason to believe it should be true.

119. Alexander A. Reinert, The Costs of Heightened Pleading, 86 Ind. L.J. 119 (2011) is an exception, but only because he considers only cases from the pre-Twombly period, and in this sense his is a completely different approach.

120. Here I focus on Tables 4 and 5 of Hubbard, Testing for Procedural Change, supra note 40, at 54 and 55, respectively (these tables report results from linear regressions, which can be viewed functionally as estimated changes in the grant rate after partialing out variation due to covariates included in Hubbard’s models). Tables 6 and 7 of his paper, at 56 and 57, respectively, use a different denominator—all filed cases, rather than all cases in which a Rule 12(b)(6) motion was filed. The resulting outcome variable, which equals the ratio of the number of measured Rule 12(b)(6) grants divided by the number of all cases filed, can be written as the product of (i) the conventional grant rate discussed here and (ii) the Rule 12(b)(6) filing rate among all filed cases. As discussed in supra note 114, I show in Jonah B. Gelbach, Selection in Motion, supra note 17, that both the presence and absence of judicial behavior effects is consistent with any combination of (i) change in the grant rate and (ii) change in the share of filed cases that face Rule 12(b)(6) motions. This means that, like his Table 4 and 5 results, Hubbard’s Table 6 and 7 results are consistent with both the presence and the absence of changes in judicial behavior as a result of Twombly and Iqbal.

121. For example, if there is enough settlement selection in the form of MS cases, B_{IR} might exceed B_{H} even with no plaintiff selection; if there is enough defendant selection or
Thus, Hubbard’s results for the change in the grant rate are consistent with the presence or absence of judicial behavior effects among MM cases (i.e., non-selection cases that face Rule 12(b)(6) motions). In terms of its claim to identification, then, Hubbard’s approach has no more payoff than any of the other grant rate-comparing approaches in the literature. This conclusion shows just how important it is to embed one’s analysis in a clear behavioral framework. Hubbard does provide a high-level discussion of litigation selection models, but he fails to make that discussion sufficiently operational to recognize that controlling for plaintiff selection alone is insufficient to identify any effect of interest. As with the rest of the empirical Twombly/Iqbal literature, Hubbard’s estimates identify a meaningful parameter only under the implausible assumption that there is neither defendant selection nor settlement selection.

C. Mapping Potential Outcomes into a Measure of Negatively Affected Cases

The foregoing discussion shows that judicial behavior effects cannot be identified without assumptions that no one should be willing to make. In Locking the Doors I showed how, even so, we can learn something of real empirical interest without making such assumptions. I reprise that demonstration in this section.

As discussed in Part III.B, the total number of cases that would have a motion to dismiss granted under Twombly/Iqbal equals the sum of the numbers of $M_D M_G$, $M_G M_D$, $A M_G$, and $S M_G$ disputes. Of these four types of disputes, the last three involve plaintiffs that are negatively affected by a change from Conley to Twombly/Iqbal, as I define the concept of “negatively affected” in Locking the Doors: the plaintiff doesn’t get to discovery under Twombly/Iqbal, while she would either get to discovery or receive a settlement under Conley. The set of negatively affected cases on which I focus attention equals the total number of $M_D M_G$, $A M_G$, and $S M_G$ disputes.122 If we could observe this

settled selection in the form of SM cases, the opposite will occur; see supra note 111. Similar points can be made about the sign of the $B_3$ term.

122. I emphasize that these are not the only disputes with negatively affected plaintiffs, given my definition of “negatively affected.” For example, plaintiffs in $A M_G$ and $S M_G$ disputes have to litigate Rule 12(b)(6) motions under Twombly/Iqbal that they would not face under Conley. See also Cecil, Waves, supra, at 13 (stating that “a plaintiff also may be ‘negatively affected’ by having to respond to a motion to dismiss, even if the motion is denied”).

Observe that the number of $A M_G$ and $S M_G$ disputes equals the number of disputes with Rule 12(b)(6) motions denied post-Twombly/Iqbal minus the number of $M_G M_D$ disputes. Meanwhile, the number of disputes with a Rule 12(b)(6) motions denied under Conley is at least as great as the number of $M_D M_G$ disputes. It follows that the number of $A M_G$ and $S M_G$ disputes is at least as great as the change in the number of Rule 12(b)(6) denials. Table 1 of the FJC INITIAL REPORT, see supra note 40, at 9, reports that Rule 12(b)(6) motions were filed in 6.9% of 3,795, or 262, pre-Twombly employment discrimination cases and in 9.0% of 3,871, or 348, post-Iqbal cases. These statistics together imply an increase of 86 Rule 12(b)(6) filings in employment discrimination cases. Table A-1 of the FJC UPDATED REPORT, see supra note 33, at 7, indicates that respondents prevailed in roughly 39% of the
ber, we could determine the fraction of negatively affected cases among all

123

share is observable, because the total number of

cases that would have Rule 12(b)(6) motions filed *if* Twombly/Iqbal were

the operative pleading standard is necessarily the total number of cases that *do*
have Rule 12(b)(6) motions filed under *Twombly/Iqbal when* that pleading

standard governs. Thus, the negatively affected share in question may be writ-

ten

\[
NAS = \frac{M_G M_{AM} + AM_G + SM_G}{MM + AM + SM}
\]

We can observe the number of cases in which a Rule 12(b)(6) motion is

granted under *Twombly/Iqbal*, but this number includes both cases in which

plaintiffs are negatively affected—the $M_G M_{AM}$, $AM_G$, and $SM_G$ cases discussed

above—and cases in which plaintiffs are unaffected. The latter category is

composed of $M_G M_{AM}$ disputes—those that would have a motion to dismiss

filed and granted under either pleading standard. Because it is impossible to observe

the same case adjudicated at the same time by the same court under different

pleading standards, we can never directly observe more than one potential out-

come of any case.123 Unfortunately, cases do not come with convenient labels

indicating what would have happened had they faced a different set of legal

rules. Thus we have no way to tell which cases with Rule 12(b)(6) motions

granted under *Twombly/Iqbal* also would have had such motions granted under

Conley—which cases, in other words, are $M_G M_{AM}$ cases. In terms of Figure 8,

we would like to know the number of disputes that occupy the middle three

cells of the final column—those that involve Rule 12(b)(6) grants that occur

under *Twombly/Iqbal* but would not occur under Conley—but all we can ob-

serve directly is the number of disputes that occupy all *four* cells of this col-

umn, including the $M_G M_{AM}$ cell.

Rule 12(b)(6) motions that the FJC coded in both the pre-*Twombly* and the post-*Iqbal*

periods. Using a denial rate of 39% in each of the pre-*Twombly* and post-*Iqbal* periods implies

that the number of AMs and SMs disputes must have increased by at least 34. This figure

implies that *Twombly* and *Iqbal* caused what might be called negative litigation effects on

plaintiffs—forcing them to litigate motions they would win but would not face under *Con-

ley*—in at least 13% (34 out of 262) of employment discrimination cases in which a Rule

12(b)(6) motion was filed in the post-*Iqbal* period. Thus, negative litigation effects might

well have been nontrivial.

123 The problem that causal effects are defined in terms of multiple potential out-

comes, whereas at most one potential outcome can be observed for any unit of interest, is

sometimes called the fundamental evaluation problem; see Gelbach, *Locking the Doors*, su-

pra note 17, at 2296. It is an unavoidable fact that this problem can be solved *only* via as-

sumptions.
Here is where the lower bound part comes into play. Even though there is no way to count the exact number of $M_{C1}M_{G}$ cases, there is a way to identify a number that must equal at least the number of $M_{C1}M_{G}$ cases. This is true because the number of cases that would have Rule 12(b)(6) motions granted under the Conley pleading standard necessarily equals the total number of cases of...
types $M_G M_G$, $M_G D$, and $M_G S$, and there can never be negative numbers of $M_G D$ or $M_G S$ cases. The total number of cases that would have Rule 12(b)(6) motions granted if Conley governed must exceed the total number of $M_G M_G$ cases. And the total number of cases that actually have motions granted under Conley is observable, because it is the number of cases that actually have motions granted when Conley governed. In terms of Figure 8, the number of disputes with Rule 12(b)(6) motions granted under Conley is the number of disputes occupying any of the cells in the bottom row. This number includes the number of disputes in the $M_G M_G$ cell, as well as any disputes in the $M_G D$ and $M_G S$ cells. Thus, the number of cases observed with Rule 12(b)(6) grants under Conley must equal at least the number of disputes in the $M_G M_G$ cell.

Consequently, subtracting the number of cases with motions granted under Conley from the number with motions granted under Twombly/Iqbal identifies a floor beneath—a known as a lower bound on—the number of cases in which plaintiffs are negatively affected, among those that face Rule 12(b)(6) motions. That is, the actual number of negatively affected cases must always be at least as great as this lower bound. Since the number of cases with Rule 12(b)(6) motions filed under Twombly/Iqbal is observable, this means that observable data can be used to determine a lower bound on the negatively affected share. To do so, one need only take the increase in the number of cases with a Rule 12(b)(6) motion granted and divide it by the number of cases in which a Rule 12(b)(6) motion is filed:

$$\text{(2) Lower bound on Negatively Affect Share for post-Iqbal Cases with 12(b)(6) Motions =}$$

$$\frac{\text{Increase in Grants}}{\text{Number of MotionsFiled after Twombly/Iqbal}}$$

124. Recall from section B, supra, that the existence of MA suits can be ruled out; this conclusion eliminates the possibility of observing any $M_A$ disputes. The same is true of $M_G M_G$ cases.

125. One exception to this statement relates to the nature of the FJC authors’ coding in the updated report. See Part IV.D, infra, on Type Z disputes.

126. See section D, infra, for a discussion of the empirical counterpart to “granted” motions as I have used that term in the present discussion.

127. A point I did not discuss in Locking the Doors is that there is also an identifiable upper bound. Since there can never be a negative number of $M_G M_G$ disputes, the total number of cases with Rule 12(b)(6) grants under Twombly/Iqbal is an upper bound on the number of negatively affected cases defined above. When we divide the former number by the number of Rule 12(b)(6) motions filed under Twombly/Iqbal, the result is the observed grant rate under Twombly/Iqbal. Thus, the observed post-Iqbal grant rate is itself an upper bound on the share of cases with negatively affected plaintiffs, as I have defined that term. Measures of the grant rate are relatively high—typically in the 60% range—so this upper bound is not very informative. Thus the identifiability of this upper bound appears to have primarily academic significance in the Twombly/Iqbal context.
At this point it is useful to pause to connect this discussion back to the role of behavioral assumptions. The behavioral framework introduced in section A, supra, was strong enough to allow us to rule out several dispute types. But even if disputes of these types were not eliminated—i.e., even if I did not impose an economic model that implies there will be no disputes of these types—the lower bound in formula (2) would still be valid.\textsuperscript{128} Ironically, then, Cecil has it exactly backward when he suggests that my empirical conclusions are “closely tethered” to important behavioral assumptions.\textsuperscript{129} In fact, the primary function of the behavioral framework here is to show the analytical shortcomings of other studies: as discussed above, unreasonably strong assumptions, such as the absence of party selection, would be necessary to justify use changes in the grant rate to measure even the judicial behavior effect among MM cases. By contrast, the power of my lower bound approach is its weak assumptions: it is empirically informative without making any assumptions at all on the pattern of party selection effects.\textsuperscript{130}

\textsuperscript{128} The seven eliminated dispute types are DS, DA, M\textsubscript{G}M, M\textsubscript{D}A, M\textsubscript{G}A, DM\textsubscript{D}, and DM\textsubscript{G}. The first two types do not involve a Rule 12(b)(6) motion under either pleading standard, so allowing them would not affect any of my analysis. Of these, only the DM\textsubscript{G} disputes would be included in the observed number of post-Iqbal grants. Plaintiffs in such cases would be negatively affected by Twombly and Iqbal, since they would spend resources litigating under Twombly/Iqbal, only to lose at the Rule 12(b)(6) stage, whereas they would avoid all litigation costs under Conley. The M\textsubscript{A} and DM\textsubscript{G} dispute types do not involve a Rule 12(b)(6) motion being granted under either pleading standard. Consequently, the presence of these disputes would not affect the numerator of formula (2). If my economic model were wrong in ruling out the presence of such case types, then, the only effect would be to enlarge the number of dispute types over which both the negatively affected share and my lower bound on it are defined: my empirical estimates would apply to a broadened set of cases, but the estimates would be correct for that set. Finally, M\textsubscript{D}M\textsubscript{A} and M\textsubscript{G}A disputes would be included in the number of cases with Rule 12(b)(6) motions granted under Conley. Thus if there were any of these dispute types, they would function only to reduce my lower bounds even further below the actual negatively affected share (to put it differently, one would need to add the number of M\textsubscript{D}M\textsubscript{A} and M\textsubscript{G}A disputes to the numerator in formula (2) just to get to the lower bound formula provided in that formula).

\textsuperscript{129} Cecil, Waves, supra note 13, at 38-39 (“I am in awe of Professor Gelbach’s skill in extending a common theoretical economic model to the pretrial litigation setting. Not being a practitioner of such dark arts, I will leave it to others to critique the economic model itself. But, I do know enough to recognize that such models are closely tethered to the assumptions that underlie their development, and I have a number of concerns about the assumptions on which Professor Gelbach erects his model. I am particularly concerned about the manner in which Professor Gelbach incorporates the findings of our studies into his economic model and the validity of his conclusion regarding the extent to which Twombly and Iqbal have restricted access to discovery and the opportunity to pursue their claims in court.”) (footnote omitted).

\textsuperscript{130} It appears that part of Cecil’s problem here is due to his mistaken belief that the model I use in Locking the Doors is an adaptation of Priest and Klein’s famous model. I address this issue further in section III.F.1, infra.
D. Empirical Implementation

As explained in Part IV.B, infra, the FJC authors collected their data on Rule 12(b)(6) motion filing and motion outcomes in different ways. Consequently, using formula (2) directly would require imputation of either the numerator or the denominator of formula (2). In Locking the Doors, I took the alternative approach of using a re-written version of formula (2). Simple if tedious algebra can be used to show that the following formula is mathematically equivalent to formula (2):

\[
(3) \text{Lower Bound in formula (2)} = \text{Change in grant rate} + (\text{Pre-Twombly grant rate}) \times m,
\]

where

\[
m = \frac{\text{Increase in number of motions filed}}{\text{Number of motions filed after Twombly/Iqbal}}
\]

This is a useful version of the lower bound formula for two reasons. First it allows me to use separate data source on grants and filings—an issue to which I shall return. Second, it allows one to see how the lower bound relates to the change in the grant rate. The lower bound formula equals that change, plus an additional term involving (i) the pre-Twombly grant rate and (ii) the relative numerosity of Rule 12(b)(6) motions before and after Twombly/Iqbal. Thus, whenever the number of Rule 12(b)(6) motions filed after Twombly/Iqbal than before, the term \(m\) will be positive, so that the lower bound must exceed the simple change in the Rule 12(b)(6) grant rate. Notice that if there were no party selection effects, then (all else equal) there would be no change in the number of motions filed, and the second term in formula (3) would equal zero. In other words, in the absence of party selection, my lower bound formula simplifies to the simple change in the grant rate.

131. Indeed, it is easy to see that in the absence of party selection effects, the negatively affected share NAS from formula (1)—and not just a lower bound on it—would equal the simple change in the grant rate. With no party selection, the only cases that would exist would be DD, SS, AA, and MM cases. Since there would be no difference in the number of Rule 12(b)(6) motions filed under the two pleading standards, the numerator of \(m\) in formula (3) would be zero, making the correction term zero. Note also that the pre-Twombly number of cases with granted Rule 12(b)(6) motions would equal the number of M_D cases, while the post-Iqbal number would equal the number of M_M cases since there would be no M_D cases (see text following note 105, supra) plus the number of M_M cases. Since the number of cases with Rule 12(b)(6) motions filed would be the same, the difference in the grant rates would equal the number of M_D cases divided by the number of MM cases and this is precisely the negatively affected share in the absence of party selection effects.

Finally, note that as an empirical matter, the converse of the claim in the text is not true. That is, finding empirical evidence that the lower bound and the change in the grant rate are equal does not generally imply the absence of selection. This is true because equality be-
As I discuss in section E, infra, though, the initial FJC report shows that the number of Rule 12(b)(6) motions filed rose substantially between 2006 and 2010. Consequently, my lower bounds on the negatively affected share will necessarily exceed the simple change in the grant rate that. Thus, the second term in formula (3) is a useful measure of the empirical importance of party selection effects.

This fact reflects an important part of my argument: when party behavior changes, there are additional routes to negative effects on cases besides the changes in judicial behavior that are incorporated in M_2M_3 cases. We should want to measure these effects. It is critical to recognize, then, that I am not arguing that the lower bound expressed in either formula (2) or formula (3) is a better measure, or a selection-corrected measure, of Hubbard’s “true effect on dismissal rates,”132 or some judicial behavior effect more generally. Rather, I am offering a different answer to the which-question-are-you-studying query133 from the one other authors in the literature have posed. Whereas other authors have sought to estimate only changes in judicial behavior, my approach is directed at learning something about both the effects of judicial behavior changes and effects due to changes in party behavior, because both types of effects impact parties. Thus, the approach set forth in Locking the Doors, and expanded on here, is sensitive to Fischman’s call for empirical researchers to increase the relevance of their research: this approach “allow[s] substantive questions to drive the[] choice of methods,” and it is “more explicit about how [it] combine[s] objective findings with contestable assumptions in order to reach normative conclusions.”134

E. Empirical Evidence: Estimates of Alternative Lower Bounds

In Table 1, I provide the data necessary to calculate the selection-related term for the three categories of cases I consider in Locking the Doors, which are civil rights; employment discrimination; and most other cases, to which I shall here refer as contract, tort and “other” cases.135 The first and second col-

---

133. See text accompanying supra note 95-96.
134. Fischman, supra note 5, at 154.
135. All data provided in this table were provided in Table 5 of Locking the Doors, at 2333, which draws data from Table 1 of the initial report, supra note 33, at 9, and Table A-1 of the updated report, supra note 33, at 7, as well as from supplemental tables generously provided to me by Joe Cecil. The civil rights category excludes ADA cases due to statutory changes in ADA law that took effect in 2008; see id., at 2288, n. 69, for details. Financial instruments cases are excluded due to concerns about the role of the financial crisis in changing the composition of these cases in unmeasurable ways; see id., at 2327. The contract, tort, and “other” cases include all case categories analyzed by the FJC reports except ADA cases, employment discrimination cases, financial instruments cases, and those in my civil rights category.
Columns present the increase in Rule 12(b)(6) motions filed between 2006-2010 and the number of such motions filed in 2010. The third column reports the ratio of the first column to the second, which yields $m$ from formula (3). This ratio is substantial for all three case types, varying from 0.17 to 0.37.

Table 1: Calculating the Selection-Related Term in Formula (3)

<table>
<thead>
<tr>
<th>Case Type</th>
<th>Rule 12(b)(6) Motions Filed</th>
<th>Ratio of first column to second</th>
<th>2006 percentage of movants prevailing</th>
<th>Second term in formula (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Civil rights</td>
<td>78</td>
<td>0.17</td>
<td>60.3</td>
<td>10.3</td>
</tr>
<tr>
<td>Employment discrimination</td>
<td>87</td>
<td>0.25</td>
<td>60.9</td>
<td>15.2</td>
</tr>
<tr>
<td>Contracts, tort and other</td>
<td>727</td>
<td>0.37</td>
<td>55.2</td>
<td>20.4</td>
</tr>
</tbody>
</table>

*Source: Table 5 of Locking the Doors, at 2333.

The fourth column of Table 1 reports the measure of the pre-<em>Twombly</em> grant rate that I used in <em>Locking the Doors</em>, which is the percentage of movants ultimately prevailing as to one or more claims in cases with adjudicated motions that were included in the updated report’s pre-<em>Twombly</em> (2006) period. The product of the third and fourth columns is the selection-related term, which I report in the final column of Table 1.

The selection-related term ranges from a low of 10.3 percentage points, for civil rights cases, to a high of 20.4 points, for contract, tort and “other” cases. These are substantial values, indicating that party selection effects are empirically important.

The remaining part of formula (3) involves only the change, between the pre-<em>Twombly</em> and post-<em>Iqbal</em> periods, in an appropriate measure of the Rule 12(b)(6) grant rate. The first two columns of Table 2 provide the pre-<em>Twombly</em> and post-<em>Iqbal</em> percentages of movants prevailing, while the third column provides the change. The fourth column repeats the selection-related terms calculated in Table 1, and the fifth column provides my lower bound.

---

136. The FJC Updated Report’s authors coded a movant as having prevailed in a case if “the court granted the last motion to dismiss in whole or in part and no opportunity to amend the complaint remained.” <em>FJC Updated Report</em>, supra note 33, at 3. For each of several case types, and in each of the pre- and post-<em>Twombly/Iqbal</em> periods for which the FJC collected data, I calculated the share of cases with Rule 12(b)(6) motions in which the movant prevailed according to this definition. The relevant data come from Table A-1 of the updated report, supra note 33, at 7 tbl.1, together with some supplemental data generously provided by Joe Cecil; the raw data used in <em>Locking the Doors</em> appear in Appendix A, Table 4 of that paper, at 2347.
Table 2: The Change in the Percentage of Movants Prevailing and the Lower bound on My Negatively Affected Share

<table>
<thead>
<tr>
<th>Category</th>
<th>Percentage of Movants Prevailing</th>
<th>Selection-related term</th>
<th>Lower bound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2006(^a)</td>
<td>2010(^b)</td>
<td>Change</td>
</tr>
<tr>
<td>Civil rights</td>
<td>60.3</td>
<td>68.1</td>
<td>7.8</td>
</tr>
<tr>
<td>Employment discrimination</td>
<td>60.9</td>
<td>61.1</td>
<td>0.2</td>
</tr>
<tr>
<td>Contracts, tort and other</td>
<td>55.2</td>
<td>56.3</td>
<td>1.1</td>
</tr>
</tbody>
</table>

\(^a\) Source: Table 4 of Locking the Doors, at 2331.

\(^b\) Source: Table 1, supra.

\(^c\) Sum of third and fourth columns.

Table 2 indicates that Twombly and Iqbal negatively affected a substantial share of post-Iqbal cases in which a Rule 12(b)(6) motion was filed. This conclusion holds for both civil rights and employment discrimination cases, for which the lower bounds are 18.1% and 15.4%, as well as the category of contract, tort and other cases, for which the lower bound is 21.5%. Importantly, less than half of the lower bounds’ magnitude comes, as an algebraic matter, from the change in the grant rate; for both employment discrimination cases and contract, tort, and “other” cases, virtually all of the lower bound is due to accounting for selection effects via the selection-related term in formula (3).

These findings indicate that the accounting for selection is a very important part of understanding Twombly and Iqbal’s effects.

F. Robustness of the Empirical Lower Bound Results

In this section, I address some criticisms that have been offered by Joe Cecil of the FJC and Professor David Engstrom in separate papers.\(^{137}\) Section 1 explains and responds to a grab-bag of reasons that Cecil has offered as reasons to reject the link between my behavioral framework and the data I use. Some of Cecil’s criticisms are prosaic, while others go to the very core of the idea of using a clearly stated behavioral framework to guide empirical study of civil litigation. As I explain below, none of Cecil’s criticisms is compelling.

Section 2 addresses Engstrom’s view (shared by Cecil) that my approach in Locking the Doors was overinclusive and should have considered only those cases in which plaintiffs lose on all—rather than one or more—of their claims as a consequence of Twombly and Iqbal. But Engstrom’s alternative lower

\(^{137}\) See Cecil, Waves, supra note 13, and Engstrom, Twiqbal Puzzle, supra note 13.
bound estimates, which are noticeably lower than those in Locking the Doors, suffer from what I call a “wrong denominator” problem that renders them non-comparable to the estimates in Locking the Doors. I show how to construct two further alternative measures—one that follows Engstrom’s suggestion that only entirely dismissed plaintiffs should be considered negatively affected, and one that follows the approach in Locking the Doors—that do not suffer from the wrong denominator problem and do allow a direct comparison. The alternative Engstrom-type measures are substantial in magnitude and, for two of the three case-type categories considered, they are statistically significant and exceed the alternative Locking the Doors-type estimates. These findings suggest the general robustness of the substance of my results to alternative units of analysis.

1. Setting fire to straw men: Priest & Klein and the 50-percent hypothesis

Cecil kicks off his discussion by stating that I adapt the model in George Priest and Benjamin Klein’s famous paper on selection in litigation. The model I sketch in Locking the Doors and the one that Priest and Klein build do share an important methodological similarity, in that they are both examples of what are sometimes called divergent expectations (“DE”) models. The key aspect of such models is that settlement is assumed to happen whenever there is positive surplus from settlement, which means whenever the parties’ subjective beliefs make it possible for a settlement to leave each party better off than she expects to be in the event of litigation. “Beliefs” here involve the parties’ subjective expectations concerning the probability the plaintiff would prevail if litigation occurred; the value of any relief the plaintiff would realize if she prevailed; and the various costs—both pecuniary and psychic—involved in litigating and negotiating a settlement.

But Cecil is wrong when he states that my model “is derived” from Priest and Klein’s. In fact, I do not “adapt” the Priest & Klein model. Priest and Klein impose considerable mathematical structure on the distribution of parties’

138. Cecil, Waves, supra note 13, at 38, n. 137 (citing George L. Priest and Benjamin Klein, The Selection of Disputes for Litigation, 13 J. LEGAL STUD. 1 (1984)).


140. Such relief could be either monetary or injunctive; what matters isn’t the form of relief but rather that the parties are capable of developing subjective beliefs concerning its cost (defendant) and value (plaintiff).

141. Cecil, Waves, supra note 13, at 44.

142. Cecil, Waves, supra note 13, at 38.
mutual beliefs concerning the plaintiff’s likelihood of winning and the actual likelihood. By contrast, I require only that the parties pursue their own interests as they understand them and that these beliefs are logically possible. These are more than academic points. Cecil attempts to tar *Locking the Doors* with the brush others have used to criticize Priest & Klein’s work, but the critiques he endorses are based alternatively (i) on assumptions that Priest & Klein make, but that I do not, or (ii) on predictions that Priest & Klein make concerning litigation selection, but that I do not.

Cecil first cites to Theodore Eisenberg for the proposition that Priest & Klein’s “model does not describe civil litigation when the issue in dispute is the extent of damages, or when an institutional litigant is worried about the effect of the outcome of the case on other litigation.” But a quick look at the part of Eisenberg’s paper to which Cecil cites shows that Eisenberg there is discussing not the presence of a selection effect generally, but rather the “clearly distinguishable . . . so-called 50 percent hypothesis.” As Eisenberg writes, this hypothesis “posits that the set of tried cases culled from the mass of underlying disputes will result in 50 percent victories for the plaintiff.”

Cecil’s second appeal in his criticism of Priest & Klein is to Steven Shavell’s well-known demonstration that, as his article is titled, *Any Frequency of*
But like the appeal to Eisenberg, this one has bite only with respect to the 50 percent hypothesis. Since the model in Locking the Doors does not rely on or endorse this hypothesis, Shavell’s results are no more problematic for my model than Eisenberg’s observations. In fact, as I elaborate in related work in progress, the model underlying Locking the Doors is consistent with a plaintiff’s win rate anywhere between 0 and 100% — just as Shavell demonstrates in his own model.150

Cecil next flags Samuel Gross and Kent Syverud’s “finding that data on outcomes of trials depart from the prediction of the Priest–Klein model.”151 Gross and Syverud’s report both that “the fifty percent hypothesis . . . thoroughly fail[s] to describe [the litigation] outcomes” that Gross and Syverud study,152 and that “our data, both for the entire set of trials and for the dominant subset of personal injury trials, are even more inconsistent with Priest and Klein’s general model as applied to cases with disputed damages than with the simple fifty percent hypothesis.”153 But neither finding is problematic for my approach, which does not impose the structural assumptions that yield the Priest & Klein predictions at issue. Cecil’s final basis for questioning my model is to cite to a paper “reviewing empirical support for and against the Priest–Klein model.”154 Once again for good measure: empirical support for and against the Priest–Klein model is beside the point.

In sum, whatever the merits of the criticisms Cecil marshals against Priest & Klein’s work on its own terms, these criticisms are non sequiturs as applied to the model I used in Locking the Doors and reprise in the present Article. I do not place the mathematical structure either on party beliefs, or on the distr-

148. Cecil, Waves, supra note 13, at 38, n. 141 (citing Steven Shavell, Any Frequency of Plaintiff Victory at Trial Is Possible, 25 J. LEGAL STUD. 493 (1996)).

149. Shavell, supra note 148, at 495 (“Although there are no errors of logic in the Priest Klein model and it is to be praised for its general and original conclusion that cases that go to trial are unrepresentative of settled cases the assumptions of the model that lead to the 50 percent tendency [for success at trial] appear to be special, and they implicitly rule out a general range of plausible situations.”).

150. See Jonah B. Gelbach, Selection in Motion: A Formal Model of Rule 12(b)(6) and the Twombly-Iqbal Shift in Pleading Policy 22 (August 29, 2012), http://ssrn.com/abstract=2138428. Among cases that actually have Rule 12(b)(6) motions filed, the grant rate can be zero, 100%, or any percentage in between.”). For more on the methodological import of Shavell’s critique of Priest and Klein, see Gelbach, supra, at 23-24, which provides a simple explanation of why DE models are consistent with any win rate between zero and one—that thus undermining the view that Shavell’s demonstration in an asymmetric information model is a per se reason to reject use of DE models.


152. Gross & Syverud, supra note 151, at 341.


bution of case quality, that Priest & Klein’s model does. None of the criticisms
from the literature that Cecil cites applies to my model, because not one of the
results demonstrated in the papers he cites is inconsistent with any prediction
my model makes.

2. Are there major problems with the model’s behavioral assumptions?

Cecil argues that economic models “are closely tethered to the assumptions
that underlie their development” and expresses concerns about a number of the
expositional assumptions that I make in sketching the model underlying Lock-
ing the Doors. He lists the following supposedly problematic assumptions:

- “Rule 12(b)(6) is the only type of MTD that can be filed”;
- “when MTDs are granted, they are always granted without leave to
  amend;”
- “each case involves a single claim”;
- “each case involves a single plaintiff and a single defendant.”

Cecil’s contention evinces a misreading of Locking the Doors and a mis-
understanding of the role these assumptions play not only in that paper’s anal-
ysis, but also, by extension, in behavioral modeling in general. As I shall dis-
cuss, the role of the assumptions Cecil criticizes is expositional, rather than
substantive: these assumptions function not to restrict the set of qualitative out-
comes that must be dealt with, but rather to sharpen and focus the methodologi-
cal discussion.

a. Other Rule 12(b) motions

As Cecil notes, Rule 12(b) does allow other bases for dismissal. But when
Cecil interprets the FJC data as he prefers to do, he does not worry that the two
FJC reports he co-authored also ignore these other motions. Of course that fact
does not imply that Cecil has made assumptions inconsistent with his data: it
simply suggests that he and his co-authors made the reasonable, simplifying as-
sumption that Rule 12(b)(6) motion practice could usefully be examined with-
out also examining behavior that, on balance, is mostly extraneous. Needless
to say, though, if ignoring other motions to dismiss is a problem for my ap-

156. Cecil, Waves, supra note 13, at 40.
158. Cecil, Waves, supra note 13, at 40.
159. Cecil, Waves, supra note 13, at 40.
160. At the outset of methodological discussion in Locking the Doors, I stated that “I
shall rely on a number of simplifying assumptions for the sake of exposition,” Locking the
Doors, supra note 17, at 2296 (emphasis added).
b. Leave to amend

As I stated in *Locking the Doors*, “[i]t is straightforward, but . . . somewhat complicated algebraically, to extend the model to account for grants with leave to amend.” 162 I stated further that when I considered an extended model that does allow for grants with leave to amend, “the main qualitative results are unaffected.” 163 As I discuss *supra*, 164 FJC data that I actually use in my empirical work codes whether plaintiffs or defendants ultimately “prevail” on claims challenged in Rule 12(b)(6) motions. As noted *supra*, 165 the FJC’s updated report explains that “[w]e identified cases in which the movant prevailed as those in which the court granted the last motion to dismiss in whole or in part and no opportunity to amend the complaint remained.” 166 My expositional assumption that motions are granted without leave to amend thus corresponds well to the FJC data I use, because both the assumption and the actual data concern the situation in which the motion to dismiss ultimately is dispositive. 167 The only exception to this correspondence involves the issue of “Type Z” cases, which I raise *de novo* in section III.G.2, *infra*.

c. Multiple claims

It is straightforward to make simple modifications to the conceptual apparatus in section A, *supra*, in order to incorporate the possibility that a plaintiff has multiple claims.

Note first that only cases in which Rule 12(b)(6) motions are granted involve negatively affected cases. To account for the possibility of multiple-
claim cases, define the outcomes of claims rather than of cases as “D”, “S”, “A”, or “M”, and similarly define M claims as M_G or M_D according to whether the defendant’s motion to dismiss as to the claim in question would be granted or denied.

Finally, observe that the advent of Twombly/Iqbal would negatively affect a plaintiff as to any given claim if such a change in pleading doctrine would cause the defendant to prevail on any of the plaintiff’s claims as a result of a Rule 12(b)(6) motion

(i) when the plaintiff would have faced and won such a motion as to that claim pre-Twombly (M_D M_G claims),

(ii) when the defendant would file an answer rather than a Rule 12(b)(6) motion as to that claim pre-Twombly, but the defendant would prevail on the motion as to the claim following Twombly/Iqbal (AM_G claims), or

(iii) when the plaintiff would secure a settlement from the defendant as to that claim pre-Twombly, but instead the claim would be terminated by a defendant’s prevailing on a Rule 12(b)(6) motion as to the claim following Twombly/Iqbal (SM_G claims).

Now define a negatively affected plaintiff as a plaintiff who is negatively affected as to any claim. With these refinements in hand, all the rest of my analysis goes through unchanged, because logic dictates that any plaintiff who is negatively affected as to a claim is negatively affected as to some claim. Thus, all that is necessary to adapt the definition of negatively affected cases to the reality of cases with multiple claims is a collection of straightforward, if word-dense, adjustments to the definitions I used in Locking the Doors.168

d. Multiple parties

The final expositional assumption that Cecil suggests is important is my assumption that each case has a single plaintiff and a single defendant. Now I will play the same game I played in the discussion of multiple claims: define the outcomes of claims by plaintiff p_i against defendant d_j as “D”, “S”, “A”, or “M”, and similarly define M claims as M_G or M_D according to whether the mo-

168. And this should not be a surprising conclusion, because Locking the Doors did not ignore the fact that the FJC-measured variables can involve multiple claims. See, e.g., Locking the Doors, supra note 17, at 2330-31 (“It is important to emphasize that the FJC codes a movant as prevailing if she prevailed on any of the claims she challenged via an initial Rule 12(b)(6) MTD”) (emphasis added); see also id. at 2331, n.177 (noting that “[t]he FJC could have instead coded movants as prevailing only if they prevailed on all claims challenged,” emphasizing this alternative approach).
tion to dismiss filed by whichever defendant filed it would be granted or denied as to the claim filed by whichever plaintiff filed it. Now define a negatively affected plaintiff as a plaintiff who is negatively affected as to some claim filed against some defendant. Once again all the rest of my analysis goes through unchanged, because logic dictates that any plaintiff who is negatively affected as to a claim filed against a defendant is negatively affected as to some claim filed against some defendant. And so all that is necessary to adapt the definition of negatively affected cases to the reality of cases with multiple parties on “each side of the v” is a collection of definitional adjustments to the simpler exposition in Locking the Doors. In sum, contra Cecil, the conceptual apparatus and the empirical implementation in Locking the Doors get on just fine.

3. The Proper Unit of Analysis: Claims, Cases, or Plaintiffs?

The empirical work in Locking the Doors and above considers a plaintiff to have been affected by a Rule 12(b)(6) grant whenever the Rule 12(b)(6) movant prevails as to one or more claims. This is the definition of “prevailing movant” adopted by Cecil and co-authors in the updated report. Both Cecil and Engstrom challenge my reliance on this approach to measuring prevailing movants. For example, Cecil writes that my “findings . . . extend[] only to cases where plaintiffs are denied the opportunity to settle or obtain access to discovery for at least one of what may be many claims in a case.” Of course that is true. And it is surely true that there are cases in which the termination of only one claim would be a minor affair. Further, given Cecil’s apparent belief that dismissals as to subsets of claims are unimportant, it is surprising that he and his co-authors collected so much data and reported so many results related to such dismissals. But it isn’t hard to imagine examples of multiple-claim cases in which a subset of claims would be quite important.

Even so, Engstrom argues that my approach suffers from overinclusiveness because

the grant-as-to-one-or-more-claims approach . . . sweeps in 12(b)(6) grants dismissing only some of the claims challenged in the motion, 12(b)(6) grants of motions that challenged only some of the plaintiff’s claims in the first place, and 12(b)(6) grants liberating purely periph-

---

169. But see the discussion of “Type Z” cases in section III.G.2, infra.
172. I noted this point in Locking the Doors, supra note 17, at 2330-31 (“It is important to emphasize that the FJC codes a movant as prevailing if she prevailed on any of the claims she challenged via an initial Rule 12(b)(6) [motion]”), immediately followed by footnote stating that “[t]he FJC could have instead coded movants as prevailing only if they prevailed on all claims challenged,” id. at 2331.
Engstrom then offers an alternative approach:

Plugging the FJC Second Study’s estimates of the post-
_Twombly_ change in the rate at which 12(b)(6) orders entirely dismissed one or
more plaintiffs from the litigation into Gelbach’s selection-accounting
framework yields a lower “negatively affected” share for all three case
types he examines, from 15.4% to 10.8% in job discrimination cases,
from 18.1% to 4.4% in civil rights cases, and from 21.5% to 11.3% amon
“Total Other” case types.174

It is important to understand what Engstrom is and isn’t claiming here.
First, Engstrom is not suggesting that I have somehow miscalculated the object
of my proposed inquiry, which is a lower bound on the share of post-
_Iqbal_ cases that faced Rule 12(b)(6) motions in which plaintiffs were negatively af-
ected on at least one claim. Instead, Engstrom is suggesting that we should be
answering a different question—namely—is a lower bound on the share of such cases in
which plaintiffs were _totally dismissed from litigation_.176 I disagree with Eng-
strom on this point, but for the sake of discussion I shall leave that disagree-
ment aside. As I shall now show, Engstrom’s preferred measure is unreliable
on its own terms, due to a missing data problem. The problem is that the data
necessary to properly calculate Engstrom’s preferred measure do not exist. The
estimates Engstrom reports are likely biased downward, possibly by substantial
magnitudes. It is possible, however, to construct alternative estimates that do
not suffer from this problem by looking at only those cases in which a plaintiff
loses at the Rule 12(b)(6) stage (thereby excluding from the denominator those
cases in which a Rule 12(b)(6) motion was filed but in which the defendant
didn’t prevail, however prevailing is measured). When I construct these esti-
mates, I find that both Engstrom’s approach and mine yield the conclusion that
_Twombly_ and _Iqbal_ negatively affected plaintiffs in a substantial share of cases
considered.

_a. The “wrong denominator” problem_

Engstrom first calculates what I shall call the “dismissed plaintiff” share,
among _all cases_ coded for the updated report:

---

173. Engstrom, _supra_ note 13, at 1227. As I shall discuss, the existence of Engstrom’s
second category here—“motions that challenged only some of the plaintiff’s claims in the
first place”—renders his own proposed alternative unworkable, due to a lack of data.
174. _Id._
175. _See supra_ text accompanying note 95.
176. Engstrom, _supra_ note 13, at 1228 (“Gelbach’s and my estimates do not differ be-
cause of something in the way the statistical analysis is performed. . . . Rather, we are meas-
uring different quantities of interest.”).
He then inserts the pre-Twombly and post-Iqbal values of this dismissed plaintiff share into formula (3), supra, using the same motion-filing data that I used to estimate $m$. Engstrom’s dismissed plaintiff share values are considerably lower than what they replace, i.e., the share of cases in which the movant prevails on one or more claims, i.e., formula (3)’s denominator. For example, in 2006, the FJC’s data show there were only forty-three civil rights cases in which a plaintiff was dismissed as a result of a Rule 12(b)(6) motion; this indicates a dismissed plaintiff share of 25.4% among all cases with a Rule 12(b)(6) motion coded. This is a much lower frequency than the 58.6% of the time that movants in these same 169 cases prevailed on one or more claims. The number Engstrom uses for his dismissed plaintiff share denominator is the total number of cases in which the updated report’s authors coded the outcome of a Rule 12(b)(6) motion. But the appropriate number would be the subset of these cases in which a plaintiff could possibly have been eliminated.

To understand the importance of this “wrong denominator” problem, consider a simple example, illustrated in Table 3. The first row of this table indicates that there are 100 cases in which Rule 12(b)(6) motions are filed. Of these 100 cases, there are sixty in which the Rule 12(b)(6) motion could not possibly eliminate a plaintiff entirely, even if all requested relief were granted (row (b)). There are another forty cases in which the motion could eliminate a plaintiff entirely (row (e)). There are three types of cases in which the movant prevails as to at least one claim: those in which the movant prevails on at least one claim, given that the motion couldn’t possibly eliminate a plaintiff entirely (row (c)); those in which the movant prevails on at least one claim but not on enough claims to eliminate a plaintiff entirely, among cases in which the motion could possibly eliminate a plaintiff entirely (row (f)); and those in which the movant could and in fact does prevail on enough claims to eliminate a plaintiff entirely (row (g)).

In terms of this example, my preferred approach both here and in *Locking the Doors* is to say that the movant prevails in 55% of cases: twenty-five row (c) cases, plus five row (f) cases, plus 25 row (g) cases, divided by 100 row (a) cases in total. An approach that considers plaintiffs to be negatively affected by *Twombly* and *Iqbal* only if they are entirely eliminated should instead determine that the movant prevails in 62.5% of cases: twenty-five row (g) cases in which the Rule 12(b)(6) motion is granted as to all of a plaintiff’s claims, divided by the forty row (e) cases in which a Rule 12(b)(6) motion actually challenges all of a plaintiff’s claims. But the approach Engstrom actually used would involve something different here. His approach amounts to (correctly)

177. FJC UPDATED REPORT, *supra* note 33, at 11.
taking the twenty-five row (g) cases, and then (incorrectly) dividing this num-
ber by the 100 row (a) cases overall, resulting in a reported dismissed plaintiff
share of 25%.
This example shows two important things. First, Engstrom’s approach can lead to substantial downward bias in measuring the frequency with which plaintiffs are entirely dismissed as a result of Rule 12(b)(6) motions, among cases in which they could be; his approach would yield a share of 25%, whereas the correct share would be 62.5%. To deal with this problem, each dismissed plaintiff share that Engstrom actually uses would have to be inflated to account for the fact that only some of the motions coded by the FJC authors are relevant to Engstrom’s desired analysis. For example, in the hypo just above, the proper inflation factor is 2.5—the ratio of the total number of motions coded to the number that Engstrom would have to use to measure what he wants to measure. In an ideal world, I would calculate the appropriate inflation factors and correct Engstrom’s calculations. But neither FJC report provides any information that could be used to measure the number of cases with Rule 12(b)(6) motions that could have eliminated a plaintiff entirely.178 Conse-

---

178. The initial report tells us how many cases had a Rule 12(b)(6) motion filed against them within ninety days of case filing. See FJC INITIAL REPORT, supra note 33, at 5. And the updated report tells us the number of cases in which the authors coded the results of orders adjudicating Rule 12(b)(6) motions. See FJC UPDATED REPORT, supra note 33, at 3. But so far as I can tell, neither report offers a single datum on the number of cases in which a Rule 12(b)(6) motion, had it been granted as to all claims the defendant challenged, could have eliminated that plaintiff entirely. This fact may be a consequence of the way the FJC
The second useful thing that Table 3’s example shows is that the share of cases in which (i) plaintiffs are entirely eliminated due to a Rule 12(b)(6) motion, among all cases in which they could possibly be so eliminated, can exceed the share of cases in which (ii) plaintiffs lose one or more claims, among all cases in which a Rule 12(b)(6) motion is filed. In other words, there is nothing inherent in my preferred approach that suggests it should lead to greater estimates than one would get with a proper application of Engstrom’s proposed approach.

b. A feasible alternative version of Engstrom’s approach yields results that buttress the qualitative conclusions in Locking the Doors

In this subpart I suggest a valid alternative that incorporates Engstrom’s suggestion to count only dismissed plaintiffs as having been negatively affected by Twombly/Iqbal. This alternative does not involve computing a lower bound on the share of negatively affected plaintiffs among those cases in which defendants filed Rule 12(b)(6) motions post-Iqbal. Instead, it involves computing a lower bound on the negatively affected share in post-Twombly/Iqbal cases whose defendants actually prevail on all claims necessary to eliminate one or more plaintiffs. To calculate this alternative measure, one must replace the number of Rule 12(b)(6) motions filed, in the denominator of equation (2), with the number of cases in which a plaintiff is eliminated. Thus, the formula for this alternative measure of the negatively affected share is:

\[
(5) \text{LB Among Eliminated Plaintiffs } = \frac{\text{Change in Number of Cases in which Plaintiff is Eliminated}}{\text{# of post-Iqbal Cases in which Plaintiff is Eliminated}}
\]

Notice that the measure in formula 5 involves only data from the updated report concerning the number of dismissed plaintiffs. Consequently, unlike Engstrom’s desired but unmeasured object of interest, formula (5) does not require knowing how many Rule 12(b)(6) motions could have eliminated one or more plaintiffs, had they been granted.180

179. In footnote 84, at 1228, Engstrom reports measures of such lower bounds that are analogous to those generated by formula (5) in the text just above. He calculates these estimates by taking the ratio of his earlier-discussed estimates to his dismissed plaintiff share (I

180. See supra text just following note 126.
In the first column of Table 4, I repeat the lower bound estimates I reported in *Locking the Doors* — that is, lower bounds calculated by implementing Engstrom’s preferred approach of counting only entirely dismissed plaintiffs as negatively affected, while using as the denominator the number of post-*Iqbal* cases in which a plaintiff is eliminated. In the table’s second column, I report estimates of still another lower bound estimand, given by formula (6):181

(6) LB Among Plaintiffs Losing on One or More Claims =

<table>
<thead>
<tr>
<th>Change in Number of Cases in which Plaintiff Loses on One or More Claims</th>
<th># of post-<em>Iqbal</em> Cases in which Plaintiff Loses on One or More Claims</th>
</tr>
</thead>
</table>

The formula (6) alternative approach uses my preferred coding of which plaintiffs are negatively affected—including those who lose on any claims, not just those who are entirely eliminated. As with formula (5), it then uses as the denominator only those plaintiffs considered as losing post-*Iqbal* (rather than all those plaintiffs who face some type of Rule 12(b)(6) motion post-*Iqbal*, as in formulas (2) and (3)). Thus, unlike the main estimates reported in Engstrom’s paper and my main estimates in *Locking the Doors*, the lower bound estimates based on formulas (5) and (6) can be used to form a meaningful comparison of the importance of the which-plaintiffs-really-lose issue that Engstrom raises.

The table shows that Engstrom’s preferred measure of the outcome of Rule 12(b)(6) motions—dismissed plaintiffs only—yields lower bounds that are substantial in magnitude in all cases. Indeed, they exceed those in *Locking the Doors* by substantial amounts for two of the three case categories, and they are statistically signficant at the 0.10 level for employment discrimination cases and at better than level 0.001 for the contract, tort, and “other” cases. The comparable estimates using the alternative approach to defining losing plaintiffs based on formula (6) are also substantial, and two of the three are statistically significant at level 0.01 or better.

In sum, both the formula (5) Engstrom approach and my preferred formula (6) alternative yield conclusions in line with the findings reported in *Locking the Doors*, did something similar in *Locking the Doors*, supra note 17, at 2335). Like my main approach in *Locking the Doors*, this approach uses formula (3)—so that it combines data from the initial report’s filing study and the updated report’s outcomes study. But otherwise, this approach functionally eliminates the problem of the wrong number of filed motions, because that wrong number appears in both the numerator and the denominator of the ratio in question here. Because it will be convenient for comparison’s sake, I shall use formula (5) instead of the approach Engstrom takes in his footnote 84.

181. Note that in terms of the example in Table 3, this formula is equivalent to one minus a ratio whose numerator is the sum of the numbers of post-*Iqbal* cases in rows (c), (f), and (g), and whose denominator is the sum of the same numbers of cases during the pre-*Twombly* period.
a substantial share of the plaintiffs considered in each case type must have been negatively affected—regardless of whether we focus on the set of all plaintiffs whose cases were challenged via Rule 12(b)(6), or on only the subset of these plaintiffs who lost in the face of such a challenge.

182. The statistical insignificance of the Engstrom-approach civil rights estimate and the Gelbach-approach employment discrimination estimate reflects both the fact that these are the smallest estimates for each of the two approaches and the fact that the numbers of cases included—the respective values of $N$, as defined in note , infra—are relatively small. This is especially clear by comparison to the $p$-values reported in the first column of Table 4, because these $p$-values are based on an estimate that incorporates information on many more cases (see Appendix B of Locking the Doors, available at http://www.yalelawjournal.org/images/documents/gelbach_appendix_b.pdf, for more).
Table 4: How Does the Choice of “Losing Plaintiff” Definition Affect the Lower Bound?

Comparing the Engstrom and Gelbach Approaches

<table>
<thead>
<tr>
<th></th>
<th>Locking the Doors Estimates</th>
<th>Formula (5) (feasible Engstrom approach)</th>
<th>Formula (6) (alternative Gelbach approach)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Civil rights</td>
<td>18.1</td>
<td>18.9</td>
<td>29.3</td>
</tr>
<tr>
<td></td>
<td>[0.004]</td>
<td>[0.180]</td>
<td>[0.005]</td>
</tr>
<tr>
<td>Employment discrimination</td>
<td>15.4</td>
<td>42.3</td>
<td>18.8</td>
</tr>
<tr>
<td></td>
<td>[0.033]</td>
<td>[0.059]</td>
<td>[0.141]</td>
</tr>
<tr>
<td>Contract, tort, and “other”</td>
<td>21.5</td>
<td>37.4</td>
<td>21.4</td>
</tr>
<tr>
<td></td>
<td>[0.000]</td>
<td>[0.000]</td>
<td>[0.000]</td>
</tr>
</tbody>
</table>

a Source: Lower bound point estimates are from Locking the Doors, Table 6, at 2334; p-values are from Appendix B, Table 2, available at http://www.yalelawjournal.org/images/documents/gelbach_appendix_b.pdf.

b Source: Author’s calculations based on data in Table B-1 of the updated report, at 11.

c Source: Author’s calculations based on data in Table A-1 of the updated report, at 7. For comparability with Engstrom’s calculations, civil rights cases include ADA-related cases; these cases are excluded from consideration in Locking the Doors.

G. Potential Confounding Factors, and a Previously Unnoted Data Issue

In responding above to specific criticisms made by Cecil and Engstrom, I mean to suggest neither that no substantive assumptions are required nor that the FJC data are perfect. In this section I first discuss the need for behavioral assumptions and then consider potential confounding factors.

183 To calculate p-values, first observe that if there are $G_{2010}$ and $G_{2006}$ randomly generated cases in which the plaintiff loses, however this outcome is defined, then the total number of cases with a losing plaintiff is $N=G_{2010}+G_{2006}$. Under the null hypothesis that no cases were negatively affected, $G_{2010}$ and $G_{2006}$ should be the same up to random error. That means the probability that a randomly drawn losing-plaintiff case comes from 2010 must equal one-half. A one-sided test of the null hypothesis that there were no negatively affected plaintiffs among those in which the plaintiff lost on a Rule 12(b)(6) motion post-Iqbal can then be based on tabulations of the binomial cumulative distribution with $N$ trials and success probability $\frac{1}{2}$. Each $p$-value reported in Table 4 is the probability that there will be no more than $G_{2010}$ successes in a $N$ Bernoulli trials with equal probability of success and failure. See Appendix B to Gelbach, Locking the Doors, supra note 17, available at http://www.yalelawjournal.org/images/documents/gelbach_appendix_b.pdf, for certain additional, and subtle, conceptual issues related to sampling error.
assumptions to rule out the possibility that other changes, unrelated to pleading standard changes, account for my empirical results. I then discuss the implications of a quirk of the FJC data that was previously unnoted.

1. Ruling Out Confounding Factors

One assumption necessary to justify the approach in Locking the Doors is that the composition of disputes that actually occur in the post-Iqbal period is not importantly different from the composition of disputes that would have occurred in this period if Twombly and Iqbal had never happened. Roughly speaking, this is equivalent to assuming that Twombly and Iqbal were the only causes of the differences in the numbers of cases and Rule 12(b)(6) motions filed, and in adjudication of those Rule 12(b)(6) motions that were filed.

I discussed this point in some detail in Locking the Doors, where I noted that if, for example, the underlying number of disputes grows at a fixed annual rate, then using the FJC data as I do would lead me to overstate the second component given in equation (3)’s lower bound formula. As I wrote in Locking the Doors, this would cause “the number of MTDs filed in the Iqbal period [to] rise simply due to the passage of time, rather than because of party selection effects.”

As I also wrote,

[p]erhaps some of the increase in the number of MTDs filed in employment discrimination and civil rights cases might come from such an exogenous growth source. But it seems very unlikely that exogenous growth in controversies can explain much of the lower bound for the total other cases nature-of-suit category. For this category, the rate at which MTDs were filed increased from 3.1% to 5.0% of case filings, which is more than a 60% increase, while the overall number of cases filed in the total other cases category actually fell slightly between the 2006 and 2010 study periods.

In addition, as I also noted in Locking the Doors, it is possible that primary behavior responded to perceived changes in the pleading standard in the period between the 2006 and 2010. Such changes could affect my results by changing the number of controversies that arise in the first place. As just one example, large employers, who are likely to be repeat-play defendants, might

184. Gelbach, Locking the Doors, supra note 17, at 2336.
186. Gelbach, Locking the Doors, supra note 17, at 2336.
expect to face reduced litigation expenses related to (either real or simply alleged) employment discrimination, as well as of non-discriminatory actions that might be perceived as discrimination. In turn, this might lead to an increase in the number of lawsuits filed by comparison to the number that would be filed post-

Iqbal were primary behavior unchanged. In the presence of such feedback effects, my results would need to be reinterpreted—of course, the same would be true of results generated by other researchers using before-and-after data.

Sometimes confounding factors can be addressed empirically using control or proxy variables. Other problematic stories have no happy statistical ending: for example, it seems likely that, as a general matter, it would be very difficult to hold primary behavior constant statistically. As I wrote in Locking the Doors, “[o]ne of the facts of life for nonexperimental empirical research . . . is that there are always such potential counterexplanations.” This is another way of stating the point that un-testable assumptions are an unavoidable part of empirical work.

2. “Type Z” Disputes

One qualification to the empirical analysis and implementation in the foregoing sections of this Part is necessary due to a quirk in the FJC data. The FJC authors’ coding scheme treats the movant as prevailing when the parties settle after a Rule 12(b)(6) motion is granted with leave to amend (such settlement could occur either before the deadline for the plaintiff to file an amended complaint, or after the plaintiff files an amended complaint but before the defendant files a new Rule 12(b)(6) motion). A preferable coding approach would have coded directly for whether a settlement occurred. It is worth asking how this feature of the data affects my results.

The typology outlined in Locking the Doors and further developed supra does not account for such a nuance. To do so, I introduce an additional potential outcome, “Z,” to capture cases in which the parties settle following a Rule 12(b)(6) motion’s grant with leave to amend, but before all possibility of

187. For example, with access to the FJC’s micro-data, one could estimate the Rule 12(b)(6) motion-filing rate in a model that includes business cycle-related variables such as the unemployment rate, retail sales data, or some other measure that might help account for macroeconomic variation.

188. Gelbach, Locking the Doors, supra note 17, at 2337.

189. Michael Abramowicz, Ian Ayres & Yair Listokin, Randomizing Law, 159 U. PENN. L. REV. 929 (2011), have recently argued that where possible, applicability of alternative legal rules should be randomized in order to facilitate measurement of policies’ effects. As has long been understood in statistics and the social sciences, random assignment can eliminate confounding factors. On the other hand (as Abramowicz, Ayres and Listokin note), randomization is not always feasible, and it does have drawbacks. For a general discussion, see James J. Heckman and Jeffrey A. Smith, Assessing the Case for Social Experiments, 9 J. ECON. PERSP. 85 (1995)). In the civil procedure context, it seems plausible that some of these drawbacks could be substantial.
amending the complaint expires. Consider a dispute that would have a Rule 12(b)(6) motion filed and denied pre-Twombly, the Rule 12(b)(6) motion would be filed and granted with leave to amend, after which the parties would settle before the deadline for the plaintiff to file an amended complaint (a Type Z dispute). To extend my taxonomy of pairs of potential outcomes under the two pleading standards, I shall refer to this dispute as having Type M\textsubscript{Z}. 

Given the possible existence of Type Z cases, the updated report’s definition of prevailing movant would include cases of Types ZZ, AZ, SZ, M\textsubscript{Z}Z, and M\textsubscript{Z}G in one or both of the periods. Because Type ZZ disputes will be coded as having movants prevail under both pleading standards, they are eliminated from my lower bound calculation when one subtracts the number of cases in which movants prevailed in the pre-Twombly period from the corresponding number in the post-Iqbal period, so they are unproblematic. And disputes of Type AZ can reasonably be viewed as involving negatively affected plaintiffs, so they also do not affect my conclusions.\textsuperscript{190}

Among post-Iqbal cases with prevailing movants, that leaves only disputes of Type SZ, M\textsubscript{Z}Z, and M\textsubscript{Z}G. If some of these disputes would not involve plaintiffs who are negatively affected by the change in the pleading standard, as I define this concept supra, then using the updated report’s data as I do could overstate the number of negatively affected plaintiffs.\textsuperscript{191} The updated report provides some limited data that can be used to get a handle on the relevance of such Type Z cases. Because the details are tedious, I relegate them to the Appendix. But the take-away point of the analysis there is that, under the scenario in which Type Z disputes make the greatest possible difference to my results,

\textsuperscript{190} Consider a Type A dispute—which is one whose defendant answers, rather than settling (or filing a Rule 12(b)(6) motion). Since the parties do not settle this dispute pre-answer/motion to dismiss, the plaintiff in such a case must expect her net gain from filing suit and litigating post-answer to be greater than cost the defendant expects from litigating post-answer. Now assume that Twombly and Iqbal affect neither the defendant’s expected post-answer cost of litigating nor the plaintiff’s expected post-answer gain from litigating (this is reasonable since Twombly/Iqbal affect only the chances of getting past the Rule 12(b)(6) stage). Now consider a Type AZ dispute. When the parties agree on a settlement under Iqbal after some Rule 12(b)(6) motion is granted, the plaintiff gives up her opportunity to file an amended complaint, and the defendant gives up her opportunity to file an answer in response to such a complaint. Thus the settlement amount must be no greater than the defendant’s expected post-answer costs of litigating under Conley. But we have seen that in a Type AZ dispute occurring under Conley, the plaintiff’s expected net gain from filing suit and litigating post-answer would exceed the defendant’s expected post-answer costs of litigating. Putting all this together establishes that the amount for which a Type AZ dispute settles under Twombly/Iqbal must be less than the plaintiff’s expected net gain from filing suit and litigating post-answer. In other words, the plaintiff in a Type AZ dispute does worse, as of the post-answer/motion to dismiss stage of litigation, under Twombly/Iqbal than under Conley.\textsuperscript{191}

To see this point, observe that formula (2)’s numerator can be written as the sum of (i) Type Z cases in which the defendant prevails post-Iqbal and (ii) the number of non-Type Z cases in which the defendant prevails, minus (iii) the number of pre-Twombly cases in which defendants prevail on Rule 12(b)(6) motions.
my lower bounds each would also fall by roughly 5 percentage points. Such changes are not trivial, but they would not have meaningfully altered the principal qualitative empirical conclusions I drew in Locking the Doors. To demonstrate, consider a slight edit of my summary of these conclusions in the introduction to Locking the Doors:

For employment discrimination and civil rights cases, switching from Conley to Twombly/Iqbal negatively affected plaintiffs in at least [10] % and at least [13] % of cases, respectively, that faced MTDs in the Iqbal period. Among cases not involving civil rights, employment discrimination, or financial instruments, Twombly/Iqbal negatively affected at least [16] % of plaintiffs facing MTDs in the Iqbal study period. These results tell us that Twombly/Iqbal negatively affected a sizable share of those plaintiffs who actually faced MTDs in the post-Iqbal period that the FJC studies.¹⁹²

The only edits to this text (besides removed footnotes) involve the replacement of the original lower bound figures with the worst-case scenario ones just discussed. Thus my basic conclusion—that “Twombly/Iqbal negatively affected a sizable share of those plaintiffs who actually faced MTDs in the post-Iqbal period” — persists with the worst-case scenario figures.

Moreover, that is the worst-case scenario, which easily might not hold. Presumably in some of the cases in which plaintiffs could have submitted an amended complaint but did not, the plaintiffs simply gave up after determining that an amended complaint was unlikely enough to make a difference.¹⁹³ And some of the cases presumably were Type ZZ cases—one whose resolution was unaffected by Twombly/Iqbal¹⁹⁴—or Type AZ cases, whose plaintiffs are negatively affected by Twombly/Iqbal. In the Appendix, I provide a set of assumptions that do not seem extreme, under which lower bound estimates that would account for Type Z disputes work out to be about 2-3 percentage points lower than the conceptually appropriate ones—roughly half the size of the worst-case effects. In sum, accounting for the presence of Type Z disputes does not alter the basic conclusions in Locking the Doors.

CONCLUSION

Empirical scholarship on civil procedure—especially involving changes in litigation rules—is often motivated by normatively important questions. But

¹⁹² Gelbach, Locking the Doors, supra note 17, at 2278-2279 (footnotes deleted).
¹⁹³ Thus, such cases would not be Type Z cases at all, since they do not involve a settlement having any value to the plaintiff—rather, they would best be considered Type Mₚ cases.
¹⁹⁴ The updated report’s coding scheme would include Type ZZ cases as “movant prevails” cases in both the 2006 and 2010 periods. So while plaintiffs in these cases are not negatively affected, these cases do not affect the properly calculated negatively affected share, because they drop out in the numerator of formula (3), supra.
empirical researchers will fail even to begin to answer many such questions unless they take seriously the fact that litigation involves human beings, who not only are motivated, but also have the ability to react to policy changes that affect the implications of their actions—in a word, agency. Consequently, it is critical to develop a behavioral framework when seeking to understand the empirical effects of developments that might alter the functioning of the civil justice system. In this paper, I have focused on these issues using the lens of Twombly, Iqbal, and contemporary debates concerning the federal civil pleading standard. Researchers in this literature have failed, writ large, to explain in behaviorally cognizable terms why they keep calculating changes in measures of the grant rate.

This paper revisits my own earlier work, showing how this continuing preoccupation with changes in grant rates is dually problematic. First, because parties can be expected to change their behavior in response to perceived changes in the pleading standard, grant rate changes do not identify judicial behavior effects (which seem to be the conceptual object of interest to researchers). Second, even if judicial behavior effects could, somehow, be identified, they are only part of the normatively relevant story of pleading standard changes. My own approach, while only a partial solution, does not suffer from these shortcomings, because it takes party selection into account.

An additional contribution of this Article comes in its response to two types of criticisms of my earlier work. Some of the criticism has been directed at the very idea of model-guided empirical research—in the realm of empirical civil procedure research, this means a rejection of the idea of taking seriously parties’ status as motivated subjects, rather than assuming they are essentially inanimate objects drawn from an urn. The basis of this criticism, though, is a misunderstanding of what is and is not involved in empirically implementing my behavioral framework. Ironically, my proposed alternative requires fewer assumptions than the grant rate-focused approach others take. A second type of criticism I address has to do with one aspect of my empirical implementation; both my methods and results survive a detailed reckoning with this criticism, too.195

I shall close this paper with the observation that empirical questions in civil procedure are too important to be answered as if people ignored changes in important incentives when choosing when and how to litigate. Parties don’t conduct their primary behavior that way, lawyers don’t plead or brief that way, and, one assumes, judges don’t decide cases that way. So we ought not to study litigation that way, either. We ought to study the civil justice system as it is: composed of human beings who might well respond to incentives.

195. In Gelbach, Can We Learn Anything About Pleading Changes From Existing Data?, supra note 143, I respond in detail to other criticisms related to the empirical implementation in Locking the Doors.
The updated report explains that Rule 12(b)(6) motions were granted with leave to amend in 143 cases in the report’s 2006 period and 400 cases in its 2010 period. Figures in the updated report indicate that in 39%, or 56, of these 2006 cases, the plaintiff did not submit an amended complaint; among 2010 cases, plaintiffs did not amend in 34%, or 136, of these cases. The scenario in which the results reported in *Locking the Doors* deviate as much as possible from the correct lower bound, is the one in which all of the latter group of cases—all 136 cases that had a Rule 12(b)(6) motion granted with leave to amend in the 2010 period and in which no amended complaint was submitted—were Type SZ or Type MZ cases in which the plaintiff was not negatively affected.

Unfortunately, the updated report does not separately report by case type the numbers of cases with Rule 12(b)(6) motions granted with leave to amend and no amended complaint filed. I therefore assume that for each case type, the incidence of such cases is proportional to the total number of cases in the updated report’s Table A-1. After eliminating from consideration the imputed numbers of 2010 period cases in which a plaintiff could have submitted an amended complaint following a grant with leave to amend, but did not so submit, I find that the 2010 grant rates would be roughly 5 percentage points lower for each of the three case types I consider. Under this worst-case scenario,
my lower bounds each would also fall by this amount, yielding lower bounds of about 13% for civil rights cases, 10% for employment discrimination cases, and 16% for my contract, tort, and “other” category.

Such changes would not have meaningfully changed the principal qualitative empirical conclusions I drew in Locking the Doors. To demonstrate, consider a slight edit or my summary of these conclusions in the introduction to Locking the Doors:

For employment discrimination and civil rights cases, switching from Conley to Twombly/Iqbal negatively affected plaintiffs in at least [10] % and at least [13]% of cases, respectively, that faced MTDs in the Iqbal period. Among cases not involving civil rights, employment discrimination, or financial instruments, Twombly/Iqbal negatively affected at least [16]% of plaintiffs facing MTDs in the Iqbal study period. These results tell us that Twombly/Iqbal negatively affected a sizable share of those plaintiffs who actually faced MTDs in the post-Iqbal period that the FJC studies.

The only edits (besides removed footnotes) involve the replacement of the original lower bound figures with the worst-case scenario ones just discussed. The basic conclusion—that “Twombly/Iqbal negatively affected a sizable share of those plaintiffs who actually faced MTDs in the post-Iqbal period”—persists with the worst-case scenario figures.

And of course, all other situations are brighter than the worst case. Presumably some of the cases in question—those in which plaintiffs could have submitted an amended complaint but did not involved plaintiffs who determined that an amended complaint was unlikely enough to make a difference that they just gave up. And some of the cases presumably were Type ZZ cases—ones whose resolution was unaffected by Twombly/Iqbal—or Type AZ cases, whose plaintiffs are negatively affected by Twombly/Iqbal. As just one example, suppose that half of the 136 cases in question in the 2010 period did involve negatively affected plaintiffs. Then the 2010 grant rates based on the updated report’s coding scheme would be only 2-3 percentage points greater than the ones conceptually appropriate for use in my lower bound calcula-

---

prevail (see Appendix A, Table 4 of Locking the Doors, supra note 17, at 2347). The 2010 grant rate thus falls from 56.3% (=100%×(337/599)) to 50.9% (=100%×(337-65)/(599-65)), a drop of 5.4 percentage points.

200. Gelbach, Locking the Doors, supra note 17, at 2278-2279 (footnotes deleted).

201. Such cases would not be Type Z cases at all.

202. The updated report’s coding scheme would include Type ZZ cases as “movant prevails” cases in both the 2006 and 2010 periods, so while plaintiffs in these cases are not negatively affected, these cases do not affect the properly calculated negatively affected share.
Under the assumption that the cases in question are proportionately distributed across case type, the reduction in the 2010 grant rate would be from 67.0% to 64.8% 
(=100%×(140-\(\frac{1}{2}\)×26)/(209-\(\frac{1}{2}\)×26)), for civil rights cases, a drop of 2.2 percentage points; for employment discrimination cases, the 2010 grant rate falls from 61.1% to 58.5% 
(=100%×(69-\(\frac{1}{2}\)×14)/(113-\(\frac{1}{2}\)×14)), a drop of 2.6 percentage points; for my contract, tort and “other” cases category, the 2010 grant rate falls from 56.3% to 53.8% 
(=100%×(337-\(\frac{1}{2}\)×65)/(599-\(\frac{1}{2}\)×65)), a drop of 2.5 percentage points.