THE *TWIQBAL* PUZZLE AND EMPIRICAL STUDY OF CIVIL PROCEDURE

David Freeman Engstrom*

* Associate Professor, Stanford Law School. Thanks to Ray Brescia, Joe Cecil, Ted Eisenberg, Nora Freeman Engstrom, Jonah Gelbach, Dan Ho, David Hoffman, Bert Huang, William Hubbard, Victor Quintanilla, and Norm Spaulding for helpful feedback, and to David Hausman and Scott Ganz for terrific research assistance. All errors are mine.

INTRODUCTION

Few developments in civil procedure have caused anything like the furor that has greeted the Supreme Court's decisions in *Bell Atlantic Corp. v. Twombly*[1] and *Ashcroft v. Iqbal*[2] (hereinafter "*Twiqbal*").[3] Indeed, earlier installments in the modern transformation of pretrial practice—from the rise of summary judgment, as symbolized by the Supreme Court's 1986 *Celotex* trilogy,[4] to the serial expansion of judicial case-management powers under Rules 16 and 26 and the related spread of "managerial judging"[5]—look like blips on the scholarly radar by comparison.[6] Yet the reaction to *Twiqbal* has not just been notable for its volume or intensity. The reaction has also, to an unusual degree, tended toward the empirical. In fact, it sometimes seems as if a hundred empirical flowers have bloomed, each purporting to capture something significant about the decisions' on-the-ground impact.[7]

---

1. 550 U.S. 544 (2007).

2. 556 U.S. 662 (2009).

3. Though these decisions should by now require little introduction, the *Twiqbal* decisions replaced "notice pleading" as announced in *Conley v. Gibson*, 355 U.S. 41, 47 (1957), with a more demanding pleading standard that requires a plaintiff to show not just a legally conceivable claim for relief but a factually "plausible" one. *See Iqbal*, 556 U.S. at 679-81 (setting forth a new two-step test that requires a judge to strike all "conclusory" allegations and then determine whether the residuum of allegations makes out a "plausible" claim for relief).

4. The "*Celotex* trilogy" is: *Anderson v. Liberty Lobby, Inc.*, 477 U.S. 242 (1986); *Celotex Corp. v. Catrett*, 477 U.S. 317 (1986); and *Matsushita Electric Industrial Co. v. Zenith Radio Corp.*, 475 U.S. 574 (1986). *See, e.g.*, Arthur R. Miller, *The Pretrial Rush to Judgment: Are the "Litigation Explosion," "Liability Crisis," and Efficiency Clichés Eroding Our Day in Court and Jury Trial Commitments?*, 78 N.Y.U. L. REV. 982, 1041 (2003) (summarizing the trilogy as follows: "*Celotex* has made it easier to make the motion, and *Anderson* and *Matsushita* have increased the chances that it will be granted"). My use of the term "symbolized" here is deliberate: empirical research suggests that the steepest increase in summary judgment filings and grants came *before* the *Celotex* trilogy, not after. *See* Joe S. Cecil et al., *A Quarter-Century of Summary Judgment Practice in Six Federal District Courts*, 4 J. EMPIRICAL LEGAL STUD. 861, 882 (2007) (analyzing data from 1975 to 2000 and concluding that the rate at which summary judgments were granted increased more between 1975 and 1986 than between 1986 and 2000).

5. *See* Miller, *supra* note 4, at 1004, 1013-15 (recounting repeated overhauls of Rules 16 and 26 to affect greater judicial control over the discovery process); Judith Resnik, *Managerial Judges*, 96 HARV. L. REV. 374, 414-31 (1982) (offering the classic account of a more managerial judicial role and its potential benefits and costs).

6. Dozens of commentators have strongly condemned the Court's *Twiqbal* move, while others have defended it. For a small sampling of the immense outpouring of academic commentary, see Robert G. Bone, *Plausibility Pleading Revisited and Revised: A Comment on* Ashcroft v. Iqbal, 85 NOTRE DAME L. REV. 849 (2010); Kevin M. Clermont & Stephen C. Yeazell, *Inventing Tests, Destabilizing Systems*, 95 IOWA L. REV. 821 (2010); Brian T. Fitzpatrick, Essay, Twombly *and* Iqbal *Reconsidered*, 87 NOTRE DAME L. REV. 1621 (2012); Edward A. Hartnett, *Taming* Twombly*, Even After* Iqbal, 158 U. PA. L. REV. 473 (2010); and Adam N. Steinman, *The Pleading Problem*, 62 STAN. L. REV. 1293 (2010).

7. Some twenty published and unpublished studies now offer systematic empirical analysis of *Twiqbal*'s impact. *See* JOE S. CECIL ET AL., FED. JUDICIAL CTR., MOTIONS TO

DISMISS FOR FAILURE TO STATE A CLAIM AFTER *IQBAL*: REPORT TO THE JUDICIAL CONFERENCE ADVISORY COMMITTEE ON CIVIL RULES (2011), *available at* http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/Publications/motioniqbal.pdf [hereinafter FJC FIRST STUDY]; JOE S. CECIL ET AL., FED. JUDICIAL CTR., UPDATE ON RESOLUTION OF RULE 12(B)(6) MOTIONS GRANTED WITH LEAVE TO AMEND: REPORT TO THE JUDICIAL CONFERENCE ADVISORY COMMITTEE ON CIVIL RULES (2011), *available at* http://www.fjc.gov/public/pdf.nsf/lookup/motioniqbal2.pdf/$file/motioniqbal2.pdf [hereinafter FJC SECOND STUDY]; Raymond H. Brescia, *The* Iqbal *Effect: The Impact of New Pleading Standards in Employment and Housing Discrimination Litigation*, 100 KY. L.J. 235 (2011-2012) [hereinafter Brescia, Iqbal *Effect*]; Raymond H. Brescia & Edward J. Ohanian, *The Politics of Procedure: An Empirical Analysis of Motion Practice in Civil Rights Litigation Under the New Plausibility Standard*, 46 AKRON L. REV. (forthcoming 2013), *available at* http://ssrn.com/abstract=2262068; Jill Curry & Matthew Ward, *Are* Twombly *&* Iqbal *Affecting Where Plaintiffs File? A Study Comparing Removal Rates by State*, 45 TEX. TECH L. REV. (forthcoming July 2013), *available at* http://ssrn.com/abstract=2143444; Scott Dodson, *A New Look: Dismissal Rates of Federal Civil Claims*, 96 JUDICATURE 127 (2012); Patricia W. Hatamyar, *The Tao of Pleading: Do* Twombly *and* Iqbal *Matter Empirically?*, 59 AM. U. L. REV. 553 (2010) [hereinafter Hatamyar, *Tao*]; William H.J. Hubbard, *Testing for Change in Procedural Standards, with Application to* Bell Atlantic v. Twombly, 42 J. LEGAL STUD. 35 (2013); Patricia Hatamyar Moore, *An Updated Quantitative Study of* Iqbal*'s Impact on 12(b)(6) Motions*, 46 U. RICH. L. REV. 603 (2012) [hereinafter Hatamyar Moore, *Updated Impact*]; Victor D. Quintanilla, *Beyond Common Sense: A Social Psychological Study of* Iqbal*'s Effect on Claims of Race Discrimination*, 17 MICH. J. RACE & L. 1 (2011); Joseph A. Seiner, *Pleading Disability*, 51 B.C. L. REV. 95 (2010) [hereinafter Seiner, *Pleading*]; Joseph A. Seiner, *The Trouble with* Twombly*: A Proposed Pleading Standard for Employment Discrimination Cases*, 2009 U. ILL. L. REV. 1011 [hereinafter Seiner, *Trouble*]; Jonah B. Gelbach, Note, *Locking the Doors to Discovery? Assessing the Effects of* Twombly *and* Iqbal *on Access to Discovery*, 121 YALE L.J. 2270 (2012) [hereinafter Gelbach, *Locking*]; Jonah B. Gelbach, *Material Facts in the Dispute over* Twombly *and* Iqbal*: Using Defense Summary Judgment Win Rates to Measure the Quality of Cases Affected by Heightened Pleading* (CELS Version, Nov. 2012) [hereinafter Gelbach, *Material Facts*]; Kendall W. Hannon, Note, *Much Ado About* Twombly*? A Study on the Impact of* Bell Atlantic Corp. v. Twombly *on 12(b)(6) Motions*, 83 NOTRE DAME L. REV. 1811 (2008); Morgan L.W. Hazelton, Procedural Postures: The Influence of Legal Change on Strategic Litigants and Judges (Preliminary Results) (Aug. 24, 2012) (unpublished manuscript), *available at* http://www.law.northwestern.edu/colloquium/politicaleconomy/documents/Procedural%20Postures%20-%20Hazelton.pdf; Kevin Heilenday & John M. de Figueiredo, Judicial Discretion and Civil Procedure: The Effect of Ideology on Rule 12(b)(6) Motions After *Twombly* and *Iqbal* (Oct. 10, 2011) (unpublished manuscript), *available at* http://ssrn.com/abstract=1989554; Victor Abel Pereyra & Benjamin Sunshine, Access-to-Justice v. Efficiency: An Empirical Study of Settlement Rates After *Twombly*/*Iqbal* (May 2, 2013) (unpublished manuscript), *available at* http://ssrn.com/abstract=2259766; *see also* Alexander A. Reinert, *The Costs of Heightened Pleading*, 86 IND. L.J. 119 (2011) (examining the relationship of thinly pleaded complaints to ultimate litigation success in the pre-*Twiqbal* period as a way to assess, albeit obliquely, *Twiqbal*'s likely effect). Many more studies offer empirical claims based on less systematic surveys of case law, *see, e.g.*, Elizabeth M. Schneider, *The Changing Shape of Federal Civil Pretrial Practice: The Disparate Impact on Civil Rights and Employment Discrimination Cases*, 158 U. PA. L. REV. 517, 533-36 (2010) (suggesting, based on a casual survey of published opinions, that *Twiqbal* is having a greater effect on civil rights cases), or else perform empirical analyses focused on how judges deploy the pleading standard in written opinions, *see* Colleen McNamara, Note, Iqbal *as Judicial Rorschach Test: An Empirical Study of District Court Interpretations of* Ashcroft v. Iqbal, 105 NW. U. L. REV. 401, 419-23 (2011); Martin H. Redish & Lee Epstein, *Bell Atlantic v. Twombly* and the Future of Pleading in the Federal Courts: A Normative and Empirical

Why the empirical turn? One reason is that the flimsiness of the Court's doctrinal analysis—particularly its insistence that it has not overruled *Conley*[8]—offers thin gruel for serious academic commentary of the traditional sort. Part of it, too, is that *Twiqbal* presents correspondingly rich empirical puzzles that cry out for analysis, particularly the Court's contention that trial judges can use their "judicial experience and common sense" to efficiently cull meritless cases based on allegations alone and without the benefit of discovery.[9] But perhaps most important of all, the profusion of empirical work since *Twiqbal* makes clear that quantitative empirical legal studies (or "ELS" to its practitioners[10]) is no longer the province of J.D./Ph.D. types working in specialized corners of the legal academy. Rather, the systematic collection and analysis of litigation-related data is now fully within the mainstream of what civil procedure scholars do. At risk of tautology, there is more empirical work this time around—compared to, say, the period following the Court's *Celotex* trilogy[11]—because more people are doing it.

Questions remain, however, as to the nature, role, and desirability of this empirical turn. Just how much can we learn from the recent spate of *Twiqbal* empiricism, whether about pretrial practice in particular or civil procedure more generally? Is the democratization of the ELS genre a healthy development, or is empirical inquiry better left in the hands of a few increasingly sophisticated technicians? And what lessons can we draw from the recent profusion of *Twiqbal* studies about what empirical study of civil procedure should look like going forward? This Essay uses the *Twiqbal* decisions and the empirical work they have spurred as a point of entry to consider these questions and reflect upon the contribution that ELS, now in its third decade,[12] has made (and can make) to the study of civil procedure.

Analysis (Nov. 18, 2008) (unpublished manuscript), *available at* http://ssrn.com/abstract=1581481.

8. *See* Bell Atl. Corp. v. Twombly, 550 U.S. 544, 562-63 (casting *Conley*'s "no set of facts" standard as "best forgotten as an incomplete, negative gloss on an accepted pleading standard" that "described the breadth of opportunity to prove what an adequate complaint claims, not the minimum standard of adequate pleading to govern a complaint's survival").

9. Ashcroft v. Iqbal, 556 U.S. 662, 679 (2009).

10. *See, e.g.*, Theodore Eisenberg, *The Origins, Nature, and Promise of Empirical Legal Studies and a Response to Concerns*, 2011 U. ILL. L. REV. 1713, 1715-19 (using the "ELS" coinage and describing the movement's origins).

11. *See infra* note 31 and accompanying text (showing the near absence of empirical study of the *Celotex* trilogy's effect on summary judgment practice in the five years following the decisions).

12. *See* Peter Cane & Herbert M. Kritzer, *Introduction* to THE OXFORD HANDBOOK OF EMPIRICAL LEGAL RESEARCH 1, 1 (Peter Cane & Herbert M. Kritzer eds., 2010) (noting ELS's rise beginning in the 1990s).

I. *TWIQBAL* AND ELS IN FULL FLOWER

A. *Gelbach and Boyd et al. on the* Twiqbal *Puzzle*

Two stellar contributions to the recent Conference on Empirical Legal Studies (CELS) at Stanford Law School—both focused, more or less, on the *Twiqbal* puzzle—provide a useful starting point for addressing the above questions by offering a glimpse of ELS in full flower.

In the first, Jonah Gelbach[13] offers the most ambitious and promising empirical test yet of *Twiqbal*'s impact on pretrial practice. Interestingly, Gelbach achieves this not by studying motions to dismiss directly, as his earlier work, and nearly all other *Twiqbal* empiricism, does.[14] Rather, he examines summary judgment motions before and after *Twiqbal* on the theory that, if the Court's assumption that trial judges can reliably gauge case merit in disposing of motions to dismiss holds true, then the rate at which judges subsequently grant defense-filed summary judgment motions should decline post-*Twiqbal* because cases that survive beyond the pleading stage should be more meritorious. His preliminary answer based on an ongoing analysis of job discrimination and contract cases before and after *Twiqbal*: summary judgment grant rates have not budged, thus calling into question the merits-screening capacity of trial judges armed with new dismissal powers.[15]

The joint contribution of Christina Boyd, David Hoffman, Zoran Obradovic, and Kosta Ristovski (hereinafter "Boyd et al.") takes a radically different, but no less illuminating, empirical tack.[16] In contrast to Gelbach's effort to isolate and quantify judicial merits-screening capacity, the Boyd et al. study offers a dazzling aerial view of pleading practice within the federal courts using spectral cluster analysis—a taxonomic tool developed in the hard sciences to characterize the relationships among different objects—to summarize the claim-level composition of lawsuits as plaintiffs plead them.[17] The result is a

13. *See* Gelbach, *Material Facts*, *supra* note 7.

14. *See* Gelbach, *Locking*, *supra* note 7, at 2294-301; *see also* Jonah B. Gelbach, Selection in Motion: A Formal Model of Rule 12(b)(6) and the *Twombly-Iqbal* Shift in Pleading Policy (Aug. 29, 2012) (unpublished manuscript), *available at* http://ssrn.com/abstract=2138428 (offering a formal, game-theoretic model of litigant selection and settlement dynamics in response to *Twiqbal*). *See generally* Appendix (cataloguing studies examining 12(b)(6) grant rates before and after *Twiqbal*).

15. Gelbach, *Material Facts*, *supra* note 7, at 9-10.

16. *See* Christina L. Boyd et al., *Building a Taxonomy of Litigation: Clusters of Causes of Action in Federal Complaints* (CELS Version, Nov. 2012). Boyd et al.'s study has since been published. Christina L. Boyd et al., *Building a Taxonomy of Litigation: Clusters of Causes of Action in Federal Complaints*, 10 J. EMPIRICAL LEGAL STUD. 253 (2013) [hereinafter Boyd et al., *Building a Taxonomy*].

17. *See* Boyd et al., *Building a Taxonomy*, *supra* note 16, at 261-62 (describing "data association methods" in sciences such as biology, zoology, psychiatry, and medicine and describing their occasional extension to legal analysis). *See generally* BRIAN S. EVERITT ET AL., CLUSTER ANALYSIS (5th ed. 2011) (reviewing cluster analysis methods).

wonderfully revelatory portrait of pleading practice and strategy that Boyd et al. achieve by allocating civil cases to a limited number of claim "clusters" and then mapping the relationships within and between them. Among other things, we learn that understanding litigation flows requires us to know that certain types of claims are often paired together—for instance, intellectual property claims with consumer protection claims, or breach of fiduciary duty claims with tax and securities claims.[18] Yet the exercise also reveals shifts in plaintiffs' pleading strategies over time, with direct relevance to the *Twiqbal* puzzle. Indeed, Boyd et al. offer preliminary evidence suggesting that the number of causes of action pled per case has declined significantly post-*Twombly*.[19] Thus, whatever the merits-screening capacities of trial judges deploying *Twiqbal*'s heightened pleading standard, the Court's decisions may have induced a dynamic litigant response.

B. *The Technological Flowering of ELS: Electronic Docketing and Computer Text Processing*

The sophistication and rigor of the Gelbach and Boyd et al. studies should by now be obvious. But it is also useful to step back and note some other ways in which they reflect the full flowering of ELS in the civil procedure space. Perhaps the most important is that both studies rely on electronic docket information as a data source. This has been critical to ELS's recent flowering, both in civil procedure and beyond.[20] Most obvious to anyone who regularly consumes empirical legal research, mandatory electronic docketing within the federal district courts—a process that was mostly complete by the mid-2000s on individual courts' PACER websites—has made it possible for researchers to construct something approaching a random sample of *all* filed cases of a given type.[21] In contrast to an earlier generation of empirical research on civil proce-

18. *See* Boyd et al., *Building a Taxonomy*, *supra* note 16, at 266 fig.4, 268 fig.5, 272.

19. *Id.* at 273-74 & fig.8. For another promising effort to measure a dynamic litigant response to *Twiqbal*, see Hazelton, *supra* note 7, at 18 (summarizing the results of a pilot version of a study using computerized linguistic analysis of complaints to find a limited post-*Twombly* increase in plaintiff use of causation and certainty language).

20. *See* David A. Hoffman et al., *Docketology, District Courts, and Doctrine*, 85 WASH. U. L. REV. 681, 728 (2007) (noting that the "recent availability of electronic dockets has the potential to spark a new way forward in empirical legal studies").

21. Note that I say "approaching a random sample" because the Boyd et al. study uses RECAP, a free digital archive of federal district court and bankruptcy case documents housed at Princeton University and sourced through the Public Access to Court Electronic Records (PACER) system, a much larger but fee-based repository of electronically filed federal court documents as maintained by each of the ninety-two U.S. district courts. *See* PUB. ACCESS TO COURT ELECTRONIC RECS., http://www.pacer.gov (last visited June 9, 2013). However, while PACER is increasingly comprehensive—as electronic filing is now mandatory in most districts—RECAP currently contains only one percent of PACER's documents, calling into question the representativeness of their sample. It should be noted that the reason Boyd et al. used RECAP rather than PACER is almost certainly that chief district judges are

dure and civil litigation,[22] and even some recent *Twiqbal*-focused empirical efforts,[23] Gelbach and Boyd et al. are studying the entire iceberg of federal litigation, not just its published-opinion or Westlaw-accessible tip.[24]

Yet the Gelbach and Boyd et al. studies well illustrate two further, and quite divergent, effects of electronic docketing. First, the ready availability of electronic docket materials has permitted a degree of technical sophistication in the construction and analysis of datasets—and, with it, a scale of empirical inquiry—that were unheard of a decade ago. As a concrete example, Gelbach uses a text-processing computer programming language to perform a relatively basic set of sorting and search tasks across thousands of electronic docket sheets to compile his sample of summary judgment motions.[25] But the uses for such technology can also take far more complex forms. To take just one (self-promoting) example, a scholar interested in testing a version of Marc Galanter's influential theory of the advantages enjoyed by repeat players as against one-shotters within litigation regimes[26] can use automated computer methods to "scrape" party and counsel names from thousands of electronic docket sheets and deposit them into a spreadsheet to construct a precise, rolling accounting of the litigation experience and successes of all actors within the

---

frequently not willing to grant PACER fee waivers for academic study despite the Judicial Conference's promulgation of rules expressly giving them authority to do so. For more on this outrage, see note 30, below.

22. *See* Mark A. Hall & Ronald F. Wright, *Systematic Content Analysis of Judicial Opinions*, 96 CALIF. L. REV. 63, 70, 80 n.66 (2008) (noting the rising use of Westlaw and Lexis among empirical legal scholars beginning in the 1980s and 1990s).

23. *See infra* note 40 and accompanying text.

24. Commentators have long warned of the perils of generalizing to the population of all disputes from a sample of published cases or, alternatively, the mix of published and unpublished cases available through legal research tools such as Westlaw and Lexis. *See* FJC FIRST STUDY, *supra* note 7, app. B at 37 & n.47 (comparing a PACER-drawn sample of motions to dismiss to holdings in Westlaw's "allfeds" database across three districts and finding substantial variation in the completeness of the holdings—from 87% in one to only 18% in another—and also substantial evidence that published orders were more likely to grant dismissal than unpublished orders); Brian N. Lizotte, *Publish or Perish: The Electronic Availability of Summary Judgments by Eight District Courts*, 2007 WIS. L. REV. 107, 130-37 (finding that only 40% of summary judgment cases collected using court docket records were available on Westlaw or Lexis and also finding substantial variation in publication practices and Westlaw/Lexis availability by judicial district and case outcome); Peter Siegelman & John J. Donohue III, *Studying the Iceberg from Its Tip: A Comparison of Published and Unpublished Employment Discrimination Cases*, 24 LAW & SOC'Y REV. 1133, 1144-49 (1990) (finding substantial differences in publication practices across judicial districts and case characteristics).

25. *See* Gelbach, *Material Facts*, *supra* note 7, at 5 & n.6 (noting use of the Practical Extraction and Reporting Language (Perl) to perform data management tasks). *See generally* THE PERL PROGRAMMING LANGUAGE, http://www.perl.org (last visited June 9, 2013) (providing documentation on Perl's text-processing functions and other capabilities).

26. *See* Marc Galanter, *Why the "Haves" Come Out Ahead: Speculations on the Limits of Legal Change*, 9 LAW & SOC'Y REV. 95 (1974).

system.[27] Utilizing these and other technologies, researchers can develop large-scale, remarkably detailed datasets in a matter of weeks compared to the months or years early ELS researchers spent constructing even rudimentary "docket profiles."[28]

Second, and in clear tension with increasing technical sophistication, electronic docketing has brought empirical legal research within the reach of a wider set of legal scholars. No longer are research efforts necessarily dependent upon large-scale funding to send researchers or runners to courthouses to review or collect docket materials.[29] Nowadays, any researcher with a PACER account—even if denied a statutorily provided academic fee waiver by chief district judges, a continuing embarrassment for the federal judiciary[30]—can

---

27. *See* David Freeman Engstrom, *Harnessing the Private Attorney General: Evidence from Qui Tam Litigation*, 112 COLUM. L. REV. 1244, 1286-89 (2012) (using this procedure to measure returns to specialization in qui tam lawsuits brought under the False Claims Act). As another example, researchers interested in the effect of amicus activity on Supreme Court decisions can train computers to code thousands of amicus briefs at a remarkable level of detail and with remarkable accuracy. *See* Alexandra Dunworth, Joshua Fischman & Daniel E. Ho, Policy Voting: What Amici Tell Us About Law 7 (Oct. 30, 2009) (unpublished manuscript), *available at* http://dho.stanford.edu/research/amici.pdf. For more on "automated content analysis" in the legal context, including some of its methodological benefits and costs, see Chad M. Oldfather et al. *Triangulating Judicial Responsiveness: Automated Content Analysis, Judicial Opinions, and the Methodology of Legal Scholarship*, 64 FLA. L. REV. 1189 (2012).

28. For examples of early efforts to create "docket profiles" focused on shifts in docket volume and content, see CHARLES CLARK, REPORT ON CIVIL CASES OF THE BUSINESS OF THE FEDERAL COURTS 3-4 (1934), *later published as* AM. LAW INST., A STUDY OF THE BUSINESS OF THE FEDERAL COURTS, PART II, CIVIL CASES (1934); FELIX FRANKFURTER & JAMES M. LANDIS, THE BUSINESS OF THE SUPREME COURT: A STUDY IN THE FEDERAL JUDICIAL SYSTEM (1927).

29. For a necessarily brief sketch of the history of empirical research on civil procedure and process throughout the twentieth century, see below notes 112-115 and accompanying text. For relatively recent examples of empirical research efforts in the pre-electronic-docketing era that required physical visits to often far-flung courthouses, see John J. Donohue III & Peter Siegelman, *Law and Macroeconomics: Employment Discrimination Litigation over the Business Cycle*, 66 S. CAL. L. REV. 709, 713 n.5 (1993) (noting data collection at federal records centers in seven cities); Theodore Eisenberg & Stewart Schwab, *The Reality of Constitutional Tort Litigation*, 72 CORNELL L. REV. 641, 651-52, 658-59 (1987) (noting that the authors examined paper docket records in the U.S. District Court for the Central District of California for their study of § 1983 civil rights cases).

30. The Judicial Conference of the United States has promulgated rules expressly authorizing courts to grant PACER fee waivers for academic research. *See* U.S. JUDICIAL CONFERENCE, ELECTRONIC PUBLIC ACCESS FEE SCHEDULES 3 (effective Apr. 1, 2013), *available at* http://pacer.psc.uscourts.gov/documents/epa_feesched.pdf (permitting courts to exempt "from payment of [PACER] fee[s] . . . individual researchers associated with educational institutions" in order to, among other things, "promote public access to information"). However, many chief district judges refuse to grant such waivers. At least one judge responded to the author's request for a fee waiver in connection with a research project examining *qui tam* litigation under the False Claims Act by noting the district's "long-standing policy not to grant exemptions to [PACER fees] for research." Letter from David J. Bradley, Clerk of Court, U.S. Dist. Court for the S. Dist. of Tex., to author (Feb. 3, 2010) (on file with

draw a random sample of cases from one or more jurisdictions and begin the search for regularities, all without leaving the office.

Given these technological advances, it should not be surprising to learn that *Twiqbal* empiricism dwarfs the empirical study performed in response to earlier tectonic shifts in pretrial practice. Vividly illustrating this change, the five years following the Court's 1986 *Celotex* trilogy—that is, roughly the same amount of time that has elapsed since *Twombly*—saw only a *single study* offering anything akin to an empirical accounting of changes in summary judgment practice in the 1986 decisions' wake.[31]

C. *ELS and Methodological Cross-Pollination*

It is not just the technical sophistication and data sourcing of the Gelbach and Boyd et al. studies that symbolize the full flowering of ELS; it is also the way the two types of studies fit together. Commentators often categorize empirical legal research based on the unit of analysis (cases, courts, judges, etc.), in interrogating data.[32] But pairing the Gelbach and Boyd et al. studies helps us to see other useful categorizations as well. Indeed, Gelbach's effort is essentially behavioralist in its orientation; his study seeks to draw inferences about judicial motivations and capacities via data on system outputs. The Boyd et al. study, by contrast, is both more descriptive in its aspiration and more synoptic in its

---

author). The categorical unwillingness of some district courts to support empirical research designed to improve the administration of civil justice is an embarrassment that calls out for correction by Congress or the Judicial Conference.

31. *See* JOE S. CECIL & C.R. DOUGLAS, FED. JUDICIAL CTR., SUMMARY JUDGMENT PRACTICE IN THREE DISTRICT COURTS 2-3, 10-11 (1987) (using docket sheets from three district courts to analyze filing and grant rates for summary judgment motions between 1975 and 1986, and finding little change in filings but a *decrease* in the rate at which such motions were granted). Beyond the Cecil and Douglas study, a pair of articles made empirical claims about the trilogy's effect, but only one performed, and then only in passing, a pre/post comparative study. *See* Samuel Issacharoff & George Loewenstein, *Second Thoughts About Summary Judgment*, 100 YALE L.J. 73, 91-92 (1990) (surveying, as part of a primarily theoretical article, all published opinions from the first quarter of 1988 that mentioned *Celotex*, and finding that 98 of 122 motions made by defendants were granted); Matthew W. Wallace, Comment, *Overruling Tradition: Summary Judgment in the Eleventh Circuit After 1986*, 41 MERCER L. REV. 737, 751 n.109 (1990) (noting, in a footnote of an otherwise theoretical comment, a lower Eleventh Circuit reversal rate based on a random sample of district court cases granting summary judgment in the three years after, as compared to before, *Celotex*). The first systematic effort to gauge any post-*Celotex* shift in summary judgment practice did not come until 1994. *See* Gregory A. Gordillo, Note, *Summary Judgment and Problems in Applying the* Celotex *Trilogy Standard*, 42 CLEV. ST. L. REV. 263, 278 & nn.107-08, 279 (1994) (analyzing dispositions in published opinions from 1979 to 1985 and from 1987 to 1992 in Ohio federal district courts and finding that the summary judgment grant rate for defendants increased from 53% to 69%).

32. *See, e.g.*, Herbert M. Kritzer, *Studying Disputes: Learning from the CLRP Experience*, 15 LAW & SOC'Y REV. 503, 504 (1980-1981) (noting "three basic approaches for collecting data about dispute processing" based on the "fundamental unit for sampling—the case, the institution, or the participant").

perspective. Indeed, if Gelbach is a lab technician cooking up experiments that can isolate judicial capacity, then Boyd et al. are cartographers, mapping the landscape of federal litigation from on high.[33]

Drawing this stylized contrast between the two study types helps us to see the critically important synergies between them. Indeed, research designs like Gelbach's depend on a strong assumption that key attributes of the legal environment do not vary across study periods. This most obviously includes governing law. Thus, if courts make an alteration mid-study to an element of the substantive liability standard—think here of the Supreme Court's recent tweaking in *Tellabs* of the scienter that securities plaintiffs must prove[34]—then the volume and nature of cases in the case pool will surely shift as well, confounding estimation of trial judges' merits-screening capacities. Similarly, if plaintiffs' lawyers experience an influx of cash midway through the study period as new sources of litigation funding come online, that may also affect litigation flows.[35]

Yet even limiting empirical study to a particular type of case or claim—recall here that Gelbach examines only job discrimination and contract causes of action—cannot fully inoculate research designs like Gelbach's from comparability concerns. For instance, the amount or severity of actionable misconduct by employers, tortfeasors, contracting parties, and the like can also change from one study period to the next, altering the pool of cases in each. To take an example from the job discrimination context, John Donohue and Peter Siegelman have shown that case filings vary with the business cycle and have argued that the variation is attributable not just to the fact that economic downturns produce more firings, but also because higher unemployment increases the amount in controversy by extending the period wrongfully terminated employees are without work and unable to mitigate damages.[36] This is important, for it suggests that macroeconomic fluctuations impact not just filing rates but also case stakes. Systematically larger cases could, in turn, alter the settlement calculus

33. This characterization of the two studies is not perfectly apt. After all, the Boyd et al. study, by showing a post-*Twiqbal* decrease in the average number of causes of action pled, also tells us something about litigant pleading strategies when initiating litigation.

34. Tellabs v. Makor Issues & Rights, 551 U.S. 308, 321-24 (2007) (resolving interpretive debate among the circuits about the "strong inference" standard in the Private Securities Litigation Reform Act). Another example is the Americans with Disabilities Act (ADA) Amendments Act of 2008 which, by broadening the ADA's coverage, likely altered the shape and size of the job discrimination case pool. *See* Gelbach, *Locking*, *supra* note 7, at 2325 n.167.

35. For more on the funding increasingly available to plaintiff-side lawyers, see Nora Freeman Engstrom, *Lawyer Lending: Costs and Consequences*, 63 DEPAUL L. REV. (forthcoming 2014).

36. Donohue & Siegelman, *supra* note 29, at 717-25. Another example comes in the securities fraud context, where scholars have long noted that the incentives to commit fraud vary with the business cycle. *See, e.g.*, Paul Povel et al., *Booms, Busts, and Fraud*, 20 REV. FIN. STUD. 1219, 1219-20 (2007) (rehearsing this debate).

across study periods,[37] once more threatening Gelbach's ability to make a valid, all-else-equal comparison.[38]

Of course, such concerns need not be fatal. But they do underscore the potential for fruitful cross-pollination between the Gelbach and Boyd et al. modes of inquiry. Work like Gelbach's will often need work like Boyd et al.'s, either to show that differences in the litigation environment across time periods are inconsequential or, alternatively, to construct measures that can control for such differences in deriving empirical estimates. And Boyd et al. need studies like Gelbach's to make their unglamorous cartography work policy relevant. To that extent, the two types of study constitute a fully mature, and fully symbiotic, empirical research agenda.

II. FORESTS, TREES, AND THE CHALLENGE OF ASSESSING PROCEDURAL CHANGE: THE LIMITS OF *TWIQBAL* EMPIRICISM

The above Part paints a rosy portrait of *Twiqbal*-related empirical efforts. And in many respects, the celebratory tone is deserved. The best empirical work exploring *Twiqbal*'s effects is plainly light-years ahead of anything produced in the wake of the Court's 1986 *Celotex* trilogy.[39] But the discussion to this point, while cheering on a pair of well-executed studies, has abstracted from the key question in all of this: just how much can we learn from the recent spate of *Twiqbal* empiricism, whether about *Twiqbal*'s on-the-ground effect or about pretrial practice and civil procedure more broadly? For the vast majority of post-*Twiqbal* empirical work beyond the Gelbach and Boyd et al. studies just noted, the answer, sadly, is not much.

This Part surfaces two kinds of problems with existing empirical studies focused on measuring 12(b)(6) grant rates before and after *Twiqbal*, and shows that those problems seem to matter—at times significantly—in terms of the inferences that can reasonably be drawn about *Twiqbal*'s effect. First are some basic measurement and methods concerns, including sampling bias and a

37. This, of course, depends on the assumption that higher stakes make cases harder to settle. *See* RICHARD A. POSNER, ECONOMIC ANALYSIS OF LAW 599 (7th ed. 2007) (noting the general view that larger-value cases are less likely to settle).

38. Nor do inferential threats come solely from changes within the particular litigation type under study, for rising case filings in one litigation area can also impact judicial behavior in other areas. A growing body of research documents the sensitivity of trial and appellate courts to docket caseloads. *See, e.g.*, Bert I. Huang, *Lightened Scrutiny*, 124 HARV. L. REV. 1109, 1116-33 (2011) (finding that circuit courts with suddenly heavy docket loads tend to reverse district court judgments with less frequency as compared to both other circuits with lighter loads and to other years in which the circuit was not so overloaded); Brian Sheppard, *Judging Under Pressure: A Behavioral Examination of the Relationship Between Legal Decisionmaking and Time*, 39 FLA. ST. U. L. REV. 931 (2012) (summarizing empirical literature on the effect of resource constraints on judicial outputs and offering an experimental simulation bolstering those results).

39. *See supra* note 31.

failure to include covariate controls in deriving empirical estimates. A second set of problems is broader and more damning: many or most of the studies do not generate—and, indeed, are not designed to generate—a useful, policy-analytic estimate of *Twiqbal*'s effect on plaintiffs' access to the legal system. In the end, the dispiriting reality is that existing *Twiqbal* empirical efforts offer precious little guidance to a Congress or an Advisory Committee considering revisions to the *Twiqbal* pleading standard.

A. *Measurement and Methods*

1. *Sampling bias*

Some of the problems with existing empirical efforts are apparent on the face of the Appendix's catalog of studies analyzing 12(b)(6) grant rates before and after *Twiqbal*. Most obviously, only a handful of the dozen-plus studies uses a fully random sample, with most instead relying on an array of Westlaw and Lexis searches—many keyed to an order's citation to *Conley* or one of the *Twiqbal* decisions—to identify 12(b)(6) motions practice during the "pre" and "post" time periods.[40]

The use of Westlaw or Lexis by itself raises serious concerns. As noted previously, Westlaw and Lexis generally hold far more published than unpublished orders, and there is also substantial variation across districts both in the completeness of holdings and in the available published/unpublished mix.[41] The resulting sampling bias possibilities are legion,[42] but the most worrying is

40. Five of the studies detailed in the Appendix use Westlaw searches keyed to citation to *Conley* or one of the *Twiqbal* decisions. *See* Hatamyar, *Tao*, *supra* note 7, at 584 n.200; Hatamyar Moore, *Updated Impact*, *supra* note 7, at 610 & nn.33-35; Seiner, *Pleading*, *supra* note 7, at 116 & n.180; Seiner, *Trouble*, *supra* note 7, at 1028 & nn.128-29; Hannon, *supra* note 7, at 1830-31 & n.135. Two of the studies use Westlaw or Lexis searches keyed to case citations as well as additional disjunctive searches designed to capture 12(b)(6) orders that lack the targeted case citations. *See* Brescia, Iqbal *Effect*, *supra* note 7, at 262-64 & n.123; Dodson, *supra* note 7, at 130. Two studies do not fully specify the type of Westlaw searches used. *See* Hubbard, *supra* note 7, at 50, 63 (detailing Westlaw-based data collection, as performed by prior researchers, but not specifying the precise search terms those researchers used); Quintanilla, *supra* note 7, at 31 (noting use of unspecified "[b]road Westlaw searches"). Only two studies—both of them by the Federal Judicial Center—deviate from this Westlaw- or Lexis-based approach by using PACER to draw a census of all 12(b)(6) orders from the time periods under study. *See* FJC FIRST STUDY, *supra* note 7, app. B at 36 n.43; FJC SECOND STUDY, *supra* note 7, at 1, 3. Note, however, that the FJC authors have since stated that their approach may have missed some cases, *see* FJC SECOND STUDY, *supra* note 7, at 1, such that their dataset might be best described as a "near-census" rather than the full case population across the twenty-three district courts included in the study.

41. *See supra* note 24 (summarizing research showing sampling problems with data collected using electronic legal research tools).

42. A more general version of the problem is that interjurisdictional variation in the availability of orders will result in oversampling from certain districts, creating a risk that changes in the distribution of types of cases, judges, or litigants—rather than *Twiqbal*'s

that district judges may alter their publication practices amid the doctrinal disorder that prevails when implementing significant, destabilizing decisions like the *Twiqbal* duo. In particular, we might expect that district judges attempting to limn the boundaries of a newly minted pleading standard will be more likely to reduce an order to a written, published decision when using *Twiqbal* to dismiss a case that would have survived under *Conley*.[43] A Westlaw-drawn sample that is overpopulated with published orders may thus systematically overrepresent grants in the post-*Twiqbal* period, exaggerating the observed *Twiqbal* effect. We might also expect that judges will be more likely to select for publication a full, plaintiff-excluding dismissal, both because of the higher stakes involved and also because outright denials, and even partial grants, may not be appealable, thereby lessening a judge's felt need to reduce an order to a written decision.[44] As a result, Westlaw- and Lexis-derived samples may not just exaggerate post-*Twiqbal* grant rates; they may also exhibit substantial skew *within* the grant pool, potentially overstating the extent to which courts are using *Twiqbal*'s heightened pleading standard to bounce plaintiffs from court entirely.[45]

---

heightened pleading standard—are driving observed changes in the 12(b)(6) grant rate before and after *Twiqbal*. For more discussion on using multivariate methods to control for these and other possibilities, see *infra* notes 50-53 and accompanying text.

43. The intuition here is that lower court implementation of significant Supreme Court decisions will tend to raise issues of first impression, which district judges may see as more worthy of publication. *See* Hoffman et al., *supra* note 20, at 701-05 (offering a "behavioral model" of opinion writing and publication that includes the novelty and importance of the legal question); Siegelman & Donohue, *supra* note 24, at 1149 (noting that district judges appear to be more likely to publish an opinion that "breaks novel legal ground"). For leading examples of cases exhibiting the doctrinal uncertainty that followed *Twiqbal*, see Swanson v. Citibank, N.A., 614 F.3d 400, 403-05, 407 (7th Cir. 2010) (reversing in part, over a spirited dissent, the district court's 12(b)(6) dismissal and raising questions about *Twiqbal*'s application going forward); *id.* at 407-12 (Posner, J., dissenting in part) (questioning the majority's characterization and application of *Twiqbal*); Braden v. Wal-Mart Stores, Inc., 588 F.3d 585, 598, 602 (8th Cir. 2009) (reversing the district court's 12(b)(6) dismissal of an ERISA claim and expressing concern about the role of information asymmetries in post-*Twiqbal* cases). On doctrinal uncertainty following *Twiqbal*, see Steinman, *supra* note 6, at 1305-06 (recounting the initial uncertainty in *Twombly*'s immediate wake as to the decision's reach beyond antitrust and its application to pro se cases).

44. Commentators have long argued that 12(b)(6) denials are less likely to be reduced to written, published orders because they are not appealable. *See, e.g.*, Dodson, *supra* note 7, at 134-35. But no commentary of which I am aware has made the further observation that even partial dismissals are only sometimes subject to immediate, interlocutory appeal under Rule 54, 29 U.S.C. § 1291, and the collateral order doctrine. *Compare* Hanni v. Am. Airlines, Inc., No. C 08-00732 CW, 2008 WL 5000237, at *1, *6-7 (N.D. Cal. Nov. 21, 2008) (granting in part defendant's motion to dismiss and denying plaintiff's motion for permission to pursue an interlocutory appeal), *with* Carder v. Continental Airlines, Inc., 636 F.3d 172, 174 (5th Cir. 2011) (noting that both the district court and appeals court had granted permission to appeal a partial 12(b)(6) dismissal as to a particular claim in the complaint).

45. To their credit, a number of authors of the studies in the Appendix acknowledge sampling concerns. *See, e.g.*, Hatamyar, *Tao*, *supra* note 7, at 609 (noting that study is not based on "a perfectly random sample"); Quintanilla, *supra* note 7, at 31 n.209 (noting that

Even beyond possible publication bias, a Westlaw- or Lexis-based sampling approach keyed to citations to *Conley* or one of the *Twiqbal* decisions—a method employed by fully half of the studies catalogued in the Appendix—is likely to be problematic. For instance, citation-keyed sampling is vulnerable to the concern that some trial judges in the pre-*Twiqbal* period may, in rehearsing the "no set of facts" standard, have passed over the musty, decades-old *Conley* in favor of more recent circuit court precedent.[46] If orders citing subsequent case law in preference to *Conley* are not randomly distributed among districts, judges, or case types, then a citation-keyed search approach may introduce substantial bias into the pre-*Twiqbal* sample.[47] Similarly, at least one Appendix

"decisions unavailable on commercial databases were not examined"); Seiner, *Pleading*, *supra* note 7, at 119 (noting possible "publication bias" based on samples constructed via Westlaw); Seiner, *Trouble*, *supra* note 7, at 1031 (noting reliance on Westlaw and conceding that "many decisions that did not result in a published opinion go undetected by this analysis"). In addition, several authors mount a defense of a Westlaw- or Lexis-based sampling approach. For instance, Dodson questions the conventional view that district judges are more likely to publish 12(b)(6) grants than denials by noting that his data show that unpublished orders have a *higher* grant rate than published orders both in the pre-*Twiqbal* period (75.0% for unpublished versus 65.8% for published) and post-*Twiqbal* period (77.8% for unpublished versus 74.5% for unpublished). *See* Dodson, *supra* note 7, at 132 tbl.2, 134-35. However, this higher unpublished-order grant rate could also reflect other regularities within his data, including the possibility that pro se cases—which are typically dismissed at a much higher rate than represented cases—are also more likely to generate unpublished orders. A full cross-tabulation is therefore necessary to evaluate his claim. Notice as well that Dodson's data shows that the post-*Twiqbal* rise in the 12(b)(6) grant rate among *published* orders was greater (65.8% pre-*Twiqbal* versus 74.5% post-*Twiqbal*, an increase of 8.7%) than the rise in the grant rate among *unpublished* orders after the decisions (75.0% pre-*Twiqbal* versus 77.8% post-*Twiqbal*, an increase of only 2.8%). *Id.* at 132 tbl.2. This is broadly consistent with the above suggestion that district judges struggling to flesh out *Twiqbal*'s new pleading standard may be more likely to publish orders granting 12(b)(6) motions, particularly grants in full, than orders denying 12(b)(6) motions. To that extent, Dodson's findings should deepen, not allay, concerns about publication selection bias in Westlaw-based samples. Least compelling among the efforts to rationalize Westlaw- and Lexis-drawn samples is the recurrent contention that "reported case bias" will be "equally present" before and after *Twiqbal*, permitting a meaningful all-else-equal comparison. Hannon, *supra* note 7, at 1829 (internal quotation marks omitted); *see also* Brescia, Iqbal *Effect*, *supra* note 7, at 260 n.113 (asserting that any biases introduced into the sample by using electronic databases "would exist throughout the entire time frame studied"); Seiner, *Trouble*, *supra* note 7, at 1031 (citing Hannon's claim that bias will be "equally present in both the pre- and post-*Twombly* case set" (quoting Hannon, *supra* note 7, at 1829)). This clearly begs many of the questions raised above regarding possible shifts in judicial publication practices and also interjurisdictional shifts in case types, judges, or litigants across the pre- and post-*Twiqbal* periods.

46. *See* Hubbard, *supra* note 7, at 44 n.13 (noting that judges deciding cases in the pre-*Twiqbal* period could just as easily choose to cite "any of the hundreds of thousands of more recent (and equally controlling) precedents").

47. For instance, if district courts within a given circuit tend to cite a specific circuit court decision over *Conley* in deciding job discrimination cases, then we might worry that such cases will be underrepresented in the pre-*Twiqbal* sample. This would be concerning if job discrimination cases tend to have a higher or lower 12(b)(6) grant rate relative to other case types. For more on this possibility and ways multivariate models might control for its effects, see notes 50-53 and accompanying text, below.

author has noted that a surprising number of post-*Twiqbal* orders upon 12(b)(6) motions do not cite either of the *Twiqbal* decisions.[48] Of course, this may be because the motions under consideration challenge only the *legal* sufficiency of a complaint's allegations, but raise no quarrel with their *factual* sufficiency, making *Conley* or its circuit-specific progeny the more natural citation. But this is precisely the point: the pre-*Twiqbal* sample *does* contain those orders—that is, at least the ones citing *Conley*—once more threatening the all-else-equal comparison on which valid pre- and post-*Twiqbal* comparisons depend.

Finally, and as an extension of the latter concern, citation-keyed sampling may fall prey to the classic legal empiricist's misstep of treating doctrinal structures as fixed, exogenous categories. As Scott Dodson's illuminating study notes, the proportion of 12(b)(6) dismissals grounded on *legal* sufficiency grounds appears to have decreased post-*Twiqbal*, while the proportion of 12(b)(6) dismissals grounded on *factual* sufficiency grounds appears to have increased.[49] This raises the possibility that district courts are using the *Twiqbal* standard to dismiss cases on factual sufficiency grounds that they previously stretched to dismiss on legal sufficiency grounds. Decisional hydraulics of this sort offer yet another way citation-keyed sampling may prove problematic: if courts invoke *Twiqbal* to dismiss cases they previously dismissed under *Conley* while at the same time continuing to cite only *Conley* in the shrinking set of post-*Twiqbal* orders focused solely on the legal sufficiency of the allegations, then estimates derived from the resulting case sample are likely to overstate *Twiqbal*'s effect.

2. *Covariate controls*

Of course, multivariate regression models that include variables designed to control for variation in outcomes by judicial district and case type may mitigate some of the above sampling concerns. Including "fixed effect" controls of this sort can help counter the sampling bias that may result if, to take an example introduced earlier, district courts within a given circuit tend to cite a non-*Conley* case for the "no set of facts" proposition when deciding certain case types that have a higher underlying dismissal rate.[50] These same multivariate techniques can also control for a more general concern about unobserved case heterogeneity—that is, the possibility that simple shifts across the pre- and

48. *See* Hannon, *supra* note 7, at 1830 (noting "hundreds" of 12(b)(6) orders in his post-*Twombly* case sample in which district courts appeared to rely on *Conley*'s standard without acknowledging *Twombly*).

49. *See* Dodson, *supra* note 7, at 132-33 & tbls.4-5.

50. For instance, if job discrimination cases have a higher underlying dismissal rate than other case types in the pre-*Twiqbal* period, then their exclusion from the sample because of systematic citation of a non-*Conley* case in rehearsing the "no set of facts" standard will result in a downwardly biased pre-*Twiqbal* dismissal rate, potentially exaggerating the observed *Twiqbal* effect when comparing pre- and post-*Twiqbal* grant rates.

post-*Twiqbal* periods in the distribution of case types, litigants, or judges are behind observed differences in outcomes. Here again, and continuing the job discrimination example from above, multivariate regression with covariate controls for case type and judicial district can ensure that a higher post-*Twiqbal* grant rate is not merely an artifact of more or different job discrimination filings, whether across all jurisdictions or in particular jurisdictions with a greater propensity to dismiss them.[51]

---

51. Another plausible example of why case-type controls are critical is the wave of cases involving financial instruments such as home mortgages in response to the 2008 economic downturn and housing market collapse, which began to enter the federal courts at roughly the same time as the *Twiqbal* decisions and have been shown to have a substantially higher 12(b)(6) dismissal rate. *See* FJC FIRST STUDY, *supra* note 7, at 9 tbl.1, 12 (noting a 214% increase in "financial instrument" cases, from 1524 to 4790 across the one-year pre- and post-downturn time periods under investigation); *id.* at 14 tbl.4, 18 tbl.7 (finding a post-*Iqbal* increase in grant rate in financial instrument cases of almost 45%—from 47% to 92% as to some or all claims—six times the increase found for any other case type).

Note, however, that there is good faith disagreement about whether covariate controls for *judicial district* (as opposed to case type) are necessary and, moreover, whether they might in fact be counterproductive. Covariate controls are clearly appropriate where district-level heterogeneity in grant rates exists because of idiosyncratic, non-*Twiqbal*-related shifts in case characteristics. For instance, an idiosyncratic corporate event in the post-*Twiqbal* period—perhaps a large company moves its corporate headquarters to another district, producing a substantial downsizing of its white-collar workforce in the district—could yield a large number of job discrimination filings that are high-value compared to the pre-*Twiqbal* run of cases and so are also lower probability cases relative to pre-*Twiqbal* cases under standard assumptions that the litigant filing calculus turns, at least in part, on a case's expected value. Under this scenario, a regression analysis that does not include covariate controls for judicial district would wrongly suggest a larger *Twiqbal* effect than is warranted. Covariate controls would similarly be indicated if some districts were to implement new case management practices post-*Twiqbal* that mute the decisions' effects as to all or certain case types. *See*, *e.g.*, FED. JUDICIAL CTR., PILOT PROJECT REGARDING INITIAL DISCOVERY PROTOCOLS FOR EMPLOYMENT CASES ALLEGING ADVERSE ACTION (2011), *available at* http://www.fjc.gov/public/pdf.nsf/lookup/DiscEmpl.pdf/$file/DiscEmpl.pdf (describing new pretrial procedure for job discrimination cases to be piloted by particular district judges). If, on the other hand, district-level heterogeneity exists not because of idiosyncratic, non-*Twiqbal*-related shifts in case characteristics but rather because litigants respond differently to the *Twiqbal* decisions across districts (perhaps arriving at different judgments about the likely stringency of lower-court implementation), then including covariate controls for judicial district risks controlling away an important, causally related part of the quantity of interest we are trying to measure. *See* Jonah B. Gelbach, Can the Dark Arts of the Dismal Science Shed Light on the Empirical Reality of Civil Procedure? 6-7 (2013) (unpublished manuscript) (on file with author) [hereinafter Gelbach, *Dark Arts*]; *see also infra* notes 80-84 and accompanying text. Here, the propriety of covariate controls for judicial district will turn on which of the above types of heterogeneity—case-characteristic differences or litigant-selection differences—are more plausible.

Of course, it also remains possible that including covariate controls, though methodologically appropriate, would not materially alter the results. *See* Gelbach, *Dark Arts*, at 8 (calculating "implied" marginal effects across case types for the FJC Second Study's multivariate models and finding only relatively small differences between the mean-comparison and multivariate estimates). *But see infra* Figure 1 (reporting a drop from 7.2% to 4.3% across the mean-comparison and multivariate estimates from the FJC Second Study's second set of results). Further analysis using actual data could clarify the extent to which observed

The problem is that many of the studies catalogued in the Appendix do not perform multivariate analyses at all, instead relying on simple mean comparisons that lack covariate controls and permit only weaker tests of statistical significance.[52] Worse, several authors make interpretive claims based on results obtained from simple mean comparisons even where those results' statistical significance disappears in an accompanying multivariate analysis.[53]

B. *The Elusiveness of Social Welfare*

If the above sampling and methods concerns were the only problems with the Appendix's *Twiqbal* grant-rate studies, then there might be room for

---

differences across mean-comparison and multivariate estimates reflect appropriate and inappropriate use of covariate controls or compositional changes across the pre- and post-*Twiqbal* case samples.

52. *See* Brescia, Iqbal *Effect*, *supra* note 7, at 269-72 (reporting only mean-comparison estimates); Seiner, *Pleading*, *supra* note 7, at 118 & tbl.1 (same); Seiner, *Trouble*, *supra* note 7, at 1029-31 (same); *see also* Dodson, *supra* note 7, at 132 & tbls.2-3 (noting that mean-comparison estimates "hold up" using regression analysis but omitting results other than *z*-scores); Quintanilla, *supra* note 7, at 35-40 & n.238 (reporting partial results of a logistic regression model for one segment of the analysis examining the effect of the race of the trial judge but not the remaining segments); Hannon, *supra* note 7, at 1838 (noting that a regression analysis supports the mean-comparison estimates but offering nothing to memorialize the analysis). The main way that *Twiqbal* empirical authors assess statistical significance in connection with simple mean comparisons for different case types and order outcomes (grant in full, in part, etc.) is the chi-squared test. But chi-squared significance testing is limited in its utility relative to multivariate regression, because the test can only show that at least one pair of results in a multiple-category distribution is meaningfully different. Put another way, a single dyad of results can dominate the joint finding of statistical significance in ways that are not observable to the analyst without further, pair-wise computation. *See, e.g.*, Quintanilla, *supra* note 7, at 37 n.233 (applying the necessary method).

53. *See, e.g.*, Hatamyar, *Tao*, *supra* note 7, at 618 tbl.4, 622 tbl.5, 624 (concluding that post-*Twiqbal* courts "appear to be granting 12(b)(6) motions at a significantly higher rate than they did under *Conley*" despite the fact that, in contrast to the mean-comparison analyses, only one out of four multivariate regression analyses found a statistically significant *Twiqbal* effect, and then only as to grants made *with* leave to amend as opposed to the more informative *without* leave to amend); *see also* Hatamyar Moore, *Updated Impact*, *supra* note 7, at 634 (asserting that the FJC First Study found "a statistically significant increase in grants with leave to amend in contract, civil rights, financial instruments, and 'other' cases" without noting that all of these estimates, save financial instruments cases, did not achieve significance in the accompanying multivariate analysis); FJC FIRST STUDY, *supra* note 7, at 19 (noting lack of statistical significance in multivariate models). Relatedly, this same author goes back and forth on whether simple mean comparisons and chi-squared tests are sufficient to demonstrate statistical significance or whether multiple regression techniques are the only route. *Compare* Hatamyar, *Tao*, *supra* note 7, at 596 (noting that "inferences—if any—to be drawn from the statistics should be confined to the regressions"), *with* Hatamyar Moore, *Updated Impact*, *supra* note 7, at 618 ("[W]hether statistically significant or not, the rate of grants in full without leave to amend in cases with represented plaintiffs increased from *Conley* to *Iqbal* in most major case types . . . ."), *and id.* at 648 (declaring that the FJC's reporting of results is "a bit misleading, because there was, after all, an increase; it was apparently just not statistically significant").

optimism about our ability to gauge the decisions' effects. Additional research aimed at estimating the extent of judicial publication bias or the available published/unpublished mix on Westlaw could permit us to correct for any measurement bias or at the least make an informed prediction about its direction and magnitude. Alternatively, we might simply credit the results of the more rigorous, multivariate analyses as providing a rough, composite portrait of *Twiqbal*'s ground-level consequences.

But the collected studies suffer from far larger problems as well: careful inspection of the research designs deployed in each of the Appendix's grant-rate studies reveals that surprisingly few permit anything approaching a social-welfare or other policy-analytic judgment about the decisions' on-the-ground effects. Indeed, virtually the entire body of *Twiqbal* empiricism misses the forest (e.g., a bottom-line judgment about *Twiqbal*'s effect on plaintiff access to the legal system) for various trees (e.g., isolating and measuring a "judicial behavior" response to the decisions).[54]

1. *Unit of analysis*

Perhaps the most easily grasped version of this problem is that nearly all of the Appendix's efforts to measure post-*Twiqbal* changes in 12(b)(6) grant rates perform the analysis *at the order or claim level, not the party level*.[55] As an example, one study treats an order dismissing a single claim upon a 12(b)(6) motion as a grant for data-coding purposes even where the dismissal is only partial and thus leaves intact at least one other claim challenged in the motion.[56]

---

54. This is not to suggest that measuring the judicial behavior response to the Supreme Court's *Twiqbal* mandate might not be useful. After all, legal scholars have long offered theoretical and empirical analyses of hierarchical relations among higher and lower courts. *See, e.g.*, Evan H. Caminker, *Why Must Inferior Courts Obey Superior Court Precedents?*, 46 STAN. L. REV. 817 (1994); Donald R. Songer et al., *The Hierarchy of Justice: Testing a Principal-Agent Model of Supreme Court-Circuit Court Interactions*, 38 AM. J. POL. SCI. 673 (1994). But as noted below, even if we view existing grant-rate studies as focused narrowly on judicial behavior—contrary to the stated intention of most—they *still* fall well short of a meaningful estimate of lower-court compliance because of their failure, among other reasons, to account for litigant selection and settlement effects in gauging the post-*Twiqbal* change in 12(b)(6) grant rates. *See infra* notes 79-83 and accompanying text. For more on whether finding a judicial behavior effect might nonetheless be informative as a prerequisite to there being a litigant selection effect, see *infra* note 74.

55. The overwhelming majority of the studies use the order as the unit of analysis, meaning each 12(b)(6) order occupies a single line in the dataset. *See infra* note 64 (noting the sole studies that adopt a party- or case-level unit of analysis). For a novel claim-level research design, see Dodson, *supra* note 7. Note, however, a possible problem with Dodson's sampling approach (beyond use of Westlaw): by apparently sampling cases but then using claims as the unit of analysis, *see id.* at 131 (describing the coding approach), his sample is not really random at the claim level given the likely correlation of outcomes in orders resolving multiple claims at once.

56. *See* Brescia, Iqbal *Effect*, *supra* note 7, at 268 n.140 (coding "decisions in which motions were granted, either in whole or in part, as 'dismissal granted'"). Several more of

Moreover, while most of the other studies separately code orders fully or partially granting 12(b)(6) motions, most ignore whether the underlying motion challenged all or only some of the claims the plaintiff asserted in her complaint.[57] And continuing the order-level trend, most of the studies do not adequately distinguish between 12(b)(6) grants with and without leave to amend[58] or attempt to trace what happens to plaintiffs given the opportunity to replead.[59] Putting these various pieces of the research-design puzzle together—and as reflected in the Appendix's second-from-last and next-to-last columns—a startling picture emerges: only three studies out of the roughly twenty *Twiqbal* empirical efforts measure the rise in 12(b)(6) grants that fully exclude a plaintiff from the litigation without leave to amend.[60] And only one of these

---

the studies similarly report findings that collapse together full and partial grants into a single grant rate in performing some of the analyses, but they also separately report results for full and partial grants. *See, e.g.*, Hatamyar, *Tao*, *supra* note 7, at 596. *See generally* Appendix (cataloguing the approach taken by various *Twiqbal* grant-rate studies along this dimension).

57. *See, e.g.*, Hatamyar, *Tao*, *supra* note 7, at 594 ("I only coded the count or counts that were challenged by a 12(b)(6) motion . . . ."); *see also* Seiner, *Trouble*, *supra* note 7, at 1028 (noting coding of orders as "granted, denied, or granted-in-part" but omitting mention of situation in which not all claims are challenged on 12(b)(6) grounds). Notice that the distinction between "full" and "partial" grants raises questions about how to treat omnibus orders that simultaneously decide multiple motions to dismiss in multiparty litigation. At least one of the studies makes clear that it treats these omnibus orders as a single order for coding purposes. *See* FJC FIRST STUDY, *supra* note 7, app. C at 41 ("[W]e counted *all* Rule 12(b)(6) motions resolved by a single order as resolving a single 12(b)(6) motion addressing multiple claims."). Most studies, however, are silent on the question, leaving one to wonder about the precise contours of the "full" and "partial" grant designation across the run of cases in the datasets. A related problem is how to code 12(b)(6) grants in multiplaintiff and multidefendant litigations, since some plaintiffs will, either out of preference or because of difficulties establishing personal jurisdiction over all defendants in a single court, bring multiple, separate lawsuits rather than omnibus, multiparty actions. *See* Engstrom, *supra* note 27, at 1290-91 (confronting a similar coding challenge in a study of qui tam litigation). This suggests that an ideal unit of analysis might be the party-litigation, not party-case, level.

58. *See, e.g.*, Hubbard, *supra* note 7, at 61-64 (describing coding but making no mention of coding for grants with or without leave to amend); Quintanilla, *supra* note 7, at 34 n.230 (same); Seiner, *Trouble*, *supra* note 7, at 1028 (same). Others only partially do so. For instance, the Hatamyar Moore studies code whether "full" grants were entered with or without leave to amend, but did not do so with respect to partial grants. *See* Hatamyar, *Tao*, *supra* note 7, at 596; Hatamyar Moore, *Updated Impact*, *supra* note 7, at 612. Moreover, the FJC study—which, as noted previously, treats an order resolving multiple motions to dismiss simultaneously as a single data point—appears to have coded an order granting any relief as to any plaintiff-defendant pairing with leave to amend as allowing an opportunity to amend, even if some claims were dismissed without leave to amend. *See* FJC FIRST STUDY, *supra* note 7, app. C at 43. This coding method may *under*estimate the claim dismissal rate without leave to amend.

59. As noted in note 61 below and also in the Appendix's catalog of *Twiqbal* empirical efforts, the sole study to follow entire dismissals with leave to amend is FJC SECOND STUDY, *supra* note 7, at 1, 3.

60. *See* FJC FIRST STUDY, *supra* note 7, at 17-18 (reporting results examining motions that were granted on all of the claims asserted by at least one plaintiff). Two other studies measure the rise in 12(b)(6) grants with case-terminating effect (that is, ending the case as to

follows entire dismissals entered *with* leave to amend in order to take accurate account of plaintiffs who replead.[61]

These are strange research design choices, for the results they generate provide an incomplete and likely misleading account of *Twiqbal*'s effect on access to the legal system. Most obviously, an elevated order- or claim-based grant rate does not necessarily mean that all affected plaintiffs are being knocked out of court entirely or that *Twiqbal* is placing them in a dread Catch-22 of needing discovery to get discovery.[62] Many trial judges could just as easily be using *Twiqbal*'s heightened pleading standard to winnow complaints that are larded up with overreaching or inapposite claims.[63] By focusing on orders and claims rather than parties as the unit of analysis, and by failing to track plaintiffs dismissed with leave to amend, virtually all of the empirical efforts to date risk exaggerating *Twiqbal*'s effect from an access-to-justice perspective because they cannot exclude the possibility that many post-*Twiqbal* plaintiffs are merely entering the discovery phase with a sharpened set of liability theories.[64]

---

all remaining plaintiffs). *See* Hatamyar Moore, *Updated Impact*, *supra* note 7, at 612 (noting coding for "[w]hether the case was entirely dismissed upon the grant of a 12(b)(6) motion without leave to amend, or whether some part of the case nevertheless remained pending"); Hubbard, *supra* note 7, at 55 (examining whether "cases terminated on an MTD" increased "as a share of all cases after *Twombly*"). The other Hatamyar Moore study likewise aims to capture the case-terminating effect of a 12(b)(6) grant but appears to do so only imperfectly. *See* Hatamyar, *Tao*, *supra* note 7, at 596 (noting coding of a pending case status variable as "'no' if any part of the case remained pending after the court's ruling, and as 'yes' if the grant of the 12(b)(6) motion (perhaps in conjunction with other rulings such as the grant of a summary judgment motion) resulted in the dismissal of the entire case").

61. *See* FJC SECOND STUDY, *supra* note 7, at 4 (reporting results of analysis tracking plaintiffs whose claims were dismissed with leave to amend).

62. *See* Hatamyar, *Tao*, *supra* note 7, at 600-01 (noting that existing studies cannot tell us whether dismissal with leave to amend is merely a "preliminary step" toward dismissal with prejudice); Hubbard, *supra* note 7, at 44 ("[A]ny study relying on opinions that rule on MTDs can at best quantify only the share of MTDs granted, not the overall rate at which filed cases are dismissed.").

63. *See* Hatamyar Moore, *Updated Impact*, *supra* note 7, at 614 (noting possibility that grants with leave to amend may in fact "enhance litigation efficiency by sharpening the issues in the case at an early stage").

64. Tracking 12(b)(6) motions with leave to amend would seem to be particularly important in light of the consistent finding in several of the *Twiqbal* studies that the rate of grants with leave to amend has increased much more post-*Twiqbal* than grants without leave to amend. *See, e.g.*, FJC FIRST STUDY, *supra* note 7, at 14 tbl.4; Hatamyar Moore, *Updated Impact*, *supra* note 7, at 618 tbl.2. This concern is further borne out by the FJC Second Study, which recodes a subsample of 543 cases used in the FJC First Study in an effort to take full account of cases in which plaintiffs were granted leave to amend upon a 12(b)(6) grant. Rerunning the same basic analysis performed in the earlier FJC First Study (which did not follow grants with leave to amend), this follow-up study finds a *lower* increase in the rate at which motions to dismiss were granted in full or in part and also a lower rate at which 12(b)(6) grants led to the entire dismissal of a plaintiff compared to the earlier study. *Compare* FJC SECOND STUDY, *supra* note 7, at 4, apps. A-B, *with* FJC FIRST STUDY, *supra* note 7, at 14 tbl.4, 18 tbl.7. This strongly suggests that the empirical efforts listed in the Appendix

2. *Selection and settlement*

A second, and even more significant, version of the forest-and-trees problem is that nearly all of the empirical efforts listed in the Appendix take no account of dynamic litigant responses in the shadow of *Twiqbal*'s more demanding pleading standard.[65] Post-*Twiqbal* plaintiffs may file fewer cases.[66] Defendants may file more motions to dismiss.[67] Litigants on either side of the "v." may prove more or less willing to come to the settlement table.[68] And would-be defendants may alter their primary conduct in light of the lower liability risk that comes from *Twiqbal*'s new procedural hurdle.[69] Given these various selection and settlement effects, focusing on only the visible part of 12(b)(6) motions practice makes little sense, for any shift in grant rates pre- and post-*Twiqbal* may overstate, or understate, the actual effect on litigant access to justice.[70] As a result, even the more rigorous and resource-intensive research

---

tend to overestimate *Twiqbal*'s effect on plaintiffs' access to the legal system. Of course, the "true" social welfare effect of all of this is hard to isolate. Streamlined cases might prove less resource-intensive to adjudicate. And litigants with fewer claims might be less likely to recover, which might be efficiency-enhancing or efficiency-reducing, depending on their relationship to an optimal level of deterrence of undesirable conduct. The point here—and the point made in Part II.B more generally—is that most existing *Twiqbal* empirical efforts leave us no better equipped to engage in broader, more normative debates of this sort.

65. *See* Gelbach, *Locking*, *supra* note 7, at 2276.

66. *Id.* at 2275 (referring to these as "[p]laintiff selection effects" (italics omitted)). Importantly, plaintiffs may be less likely to file both because they are more likely to suffer dismissal under *Twiqbal*'s more demanding pleading standard and also because *Twiqbal*, by increasing the likelihood that a complaint will draw a motion to dismiss, will raise the average expected cost of litigating the case, thus lowering its expected value. *See* SEAN FARHANG, THE LITIGATION STATE: PUBLIC REGULATION AND PRIVATE LAWSUITS IN THE UNITED STATES 22 (2010) (summarizing standard models of the decision to litigate as a calculation of the expected value, net of costs, of filing suit).

67. Gelbach, *Locking*, *supra* note 7, at 2275 (referring to these as "[d]efendant selection effects" (italics omitted)).

68. *Id.* at 2276 (referring to these as "[s]ettlement selection effects" (italics omitted)). For a much earlier discussion of the possible selection and settlement effects, focused on the representativeness of published opinions, see Siegelman & Donohue, *supra* note 24, at 1147-51.

69. Gelbach, *Locking*, *supra* note 7, at 2333 (noting the possibility that employers will respond to the lower risk of liability after *Twiqbal* by "engag[ing] in more discrimination, or be[ing] less vigilant in policing any unlawful behavior of supervisors").

70. *Id.* at 2311-14 (setting forth numerical examples and showing that "neither the direction nor the magnitude of the difference in [motion to dismiss] grant rates across pleading regimes tells us anything about the magnitude of any judicial behavior effects"). An alternative way of understanding this is to see how litigant perceptions and the "true" judicial response to *Twiqbal* work in tandem. For instance, if would-be plaintiffs *over*estimate the stringency with which judges are deploying the new *Twiqbal* standard, then they will file fewer and stronger claims—and, in so doing, hold back even some claims that would survive a motion to dismiss—yielding a lower 12(b)(6) grant rate than judicial behavior alone would produce. If, by contrast, plaintiffs *under*estimate the stringency of the judicial response, then they will continue to send cases into the teeth of *Twiqbal*'s heightened pleading standard, yielding a *higher* 12(b)(6) grant rate than judicial behavior alone would produce. Most ar-

efforts among the Appendix's entries tell us surprisingly little about *Twiqbal*'s overall impact on the litigation landscape in the ways that may well matter most.

To be sure, some of the more sophisticated studies grapple with selection and settlement effects. Hubbard, for instance, focuses his analysis on what we might call "straddle" cases—that is, cases filed *before* the *Twombly* decision but decided on 12(b)(6) grounds *after* it—as a way to isolate the judicial response by washing out any effect *Twombly* had on plaintiffs' filing calculus.[71] But note some problems. As an initial matter, the fact that Hubbard's straddle cases were all decided post-*Twombly* but pre-*Iqbal* is problematic in light of the previously noted uncertainty about *Twombly*'s precise contours and trans-substantive reach before *Iqbal* clarified matters.[72] As a result, the empirical identification Hubbard's ingenious approach achieves comes at the cost of measuring the judicial response to *Twiqbal*'s procedural change precisely when doctrine was at its most fluid. More broadly, Hubbard's "straddle" approach exhibits—albeit in a very sophisticated way—the same fetish as does nearly all other *Twiqbal*-related empiricism for modeling the judicial behavior response to the decisions at the expense of forming a broader judgment about their effects on litigants.[73] Indeed, by washing out *Twiqbal*'s likely chilling effect on plaintiffs' claiming behavior, Hubbard controls away the concern that has most occupied courts and commentators and may prove the most significant of the decisions' effects on the litigation landscape.[74]

---

resting of all, if plaintiff perceptions and the "true" judicial behavior response are in full alignment, then we might expect to see no change in 12(b)(6) grant rates pre- and post-*Twiqbal*, as plaintiffs will perfectly anticipate whether their complaint will fall at the motion-to-dismiss stage. A similar analysis can be applied to defendants' willingness to file more motions to dismiss, potentially biasing estimates of *Twiqbal*'s effect.

71. *See* Hubbard, *supra* note 7, at 51 ("To control for selection effects in the composition of filed cases, my empirical strategy focuses on cases that are filed under the old standard but decided under the new standard."). Other studies have noted the possible value of a straddle approach, but not enough to implement it in any systematic way. *See* Seiner, *Trouble*, *supra* note 7, at 1029 (noting that restricting analysis to the year immediately before and after *Twombly* helps provide an accurate picture of litigation trends "immediately before the decision was issued" and control for plaintiff selection effects since "many of the complaints would have been drafted prior to the opinion").

72. *See* Hubbard, *supra* note 7, at 40 (noting that the dataset comes from "published district court opinions ruling on MTDs between May 21, 2006 and May 21, 2008 (a year before and after *Twombly*)"); *supra* note 43 (noting uncertainty in *Twombly*'s wake); *see also* Hannon, *supra* note 7, at 1830 & n.136 (noting "hundreds" of district court orders in the immediate post-*Twombly* period that cite only *Conley*, and speculating that the Court's decision had not yet "trickled down" to lower courts).

73. *See* Hubbard, *supra* note 7, at 35 (noting the study's focus on "[q]uantifying change in legal standards—in the sense of change in the propensity of judges to decide cases a certain way").

74. *See, e.g.*, Arthur R. Miller, *From* Conley *to* Twombly *to* Iqbal*: A Double Play on the Federal Rules of Civil Procedure*, 60 DUKE L.J. 1, 71 (2010) (arguing that a paramount concern about *Twiqbal* is its chilling effect on plaintiff claiming behavior); *see also* Lonny Hoffman, Twombly *and* Iqbal*'s Measure: An Assessment of the Federal Judicial Center's*

A more systematic effort to corral dynamic litigant responses in *Twiqbal*'s shadow is Gelbach's earlier published work.[75] In it, Gelbach offers an impressive theoretical framework that accounts for, rather than controlling away, possible litigant responses to *Twiqbal*'s heightened pleading standard.[76] The beauty of Gelbach's framework is that it permits researchers to recover a "lower-bound"[77] estimate of the proportion of plaintiffs whose claims were subject to a 12(b)(6) motion who were "negatively affected" by *Twiqbal* by using data-driven estimates of motion-to-dismiss filing trends and the pre-*Twiqbal* grant rate to make adjustments to the observed post-*Twiqbal* change in grant rates.[78]

Study of Motions to Dismiss, 6 FED. CTS. L. REV. 1, 28 (2011) (same); *see also supra* note 19 (noting Boyd et al. and Hazelton studies showing a dynamic litigant response to the *Twiqbal* decisions); *infra* Table 1 and accompanying text (offering evidence that litigant selection effects dominate judicial behavioral effects in terms of *Twiqbal*'s overall effect on plaintiff access to the legal system). In defense of the focus of much *Twiqbal* empiricism on measuring changes in judicial behavior, one might argue that party behavioral changes will be parasitic on judicial behavioral change, such that the chilling effect on plaintiffs that concerns many commentators is likely to occur only if judges in fact change their behavior. Conversely, if lower court judges do not change their behavior at all in response to the Supreme Court's efforts to mandate a higher pleading standard, then plaintiffs are unlikely, in equilibrium, to change their behavior either. Put another way, while a change in behavior by judges may or may not affect parties' behavior, the converse is not true; no change by judges implies no effect. The problem with this view is two-fold: First, it assumes that party behavior and judicial behavior will ultimately equilibrate, which is an (as-yet-untested) empirical claim. For the moment, it seems at least plausible that litigants will chronically over- or under-estimate the post-*Twiqbal* judicial propensity to dismiss cases given the lack of reliable information within decentralized litigation regimes about case outcomes. Second, and relatedly, even if measuring judicial behavioral changes via grant-rate studies can offer a prediction about whether a procedural change is likely to induce a claim-chilling behavioral response, the results of such an analysis do not readily translate into an estimate of the magnitude of the ground-level impact of that same procedural change on litigants, whether overall or across litigation areas. To that extent, studies focused on isolating judicial behavioral effects are less valuable to would-be rule reformers.

75. *See* Gelbach, *Locking*, *supra* note 7.

76. Gelbach does so by allocating hypothetical cases to couplets reflecting all possible case outcomes under the *Twiqbal* pleading standard as compared to a counterfactual world in which the less exacting *Conley* standard remained in force. For instance, cases that were filed under *Conley* might not be filed under *Twiqbal*, creating a "filed-under-*Conley*/not-filed-under-*Twiqbal*" couplet. Cases that would not have drawn a motion to dismiss pre-*Twiqbal* might do so post-*Twiqbal*, creating an "answered-under-*Conley*/motion-to-dismiss-under-*Twiqbal*" couplet. *See id.* at 2297-98 & fig.1.

77. *See id.* at 2277 n.21 ("A lower bound on one function's value . . . is another function with the property that the second function never takes on a value greater than the value taken on by the first function."). More colloquially, a lower bound as used here means the minimum likely value without excluding the possibility that the "true" value is in fact greater.

78. *See* Gelbach, *Locking*, *supra* note 7, at 2322-24 (explaining the construction of a "correction term" to adjust for selection and settlement effects). The construction of the correction term that Gelbach's model develops is clearly explained in his article, and so I will not attempt a rehearsal here. For now, note that Gelbach's correction machine requires a researcher to add together the pre- and post-*Twiqbal* change in grant rates and a correction term that takes the following form:

And applying his framework also appears to have a powerful impact on empirical estimates of the *Twiqbal* effect: after running the results from the Federal Judicial Center (FJC) study through his selection-accounting machine, Gelbach reports that at least 15.4% of plaintiffs in job discrimination cases, 18.1% of civil rights plaintiffs, and 21.5% of plaintiffs in "Total Other" case types whose claims were subject to a 12(b)(6) motion to dismiss post-*Iqbal* were "negatively affected" by the decisions—well above the single-digit estimates of the *Twiqbal* effect found in the more rigorous among studies in the Appendix.[79]

Yet while Gelbach's methodological framework provides a model for future researchers, the empirical portion of his analysis is less sure-footed. The main reason is that Gelbach uses data and estimates from one of the FJC studies to derive his lower-bound negatively affected share estimates, raising a pair of concerns. One of these should by now have a familiar ring: Gelbach's grant-rate estimates derive from the FJC's order-level analysis that considers only whether the order granted dismissal as to one or more of the claims challenged in the motion, *not* the part of the FJC's party-level analysis that tracks whether at least one plaintiff was entirely dismissed from the litigation.[80] And this

---

(Pre-*Twiqbal* Grant Rate) ×
(Total Post-*Twiqbal* 12(b)(6) Filings − Total Pre-*Twiqbal* 12(b)(6) Filings) /
(Total Post-*Twiqbal* 12(b)(6) Filings)

See id. at 2323. Thus, deriving a selection-adjusted estimate using Gelbach's framework requires four total pieces of information: the 12(b)(6) grant rate pre- and post-*Twiqbal* and also the total number of 12(b)(6) filings pre- and post-*Twiqbal*.

79. *Id.* at 2331-32. For instance, the FJC Second Study—which, as I will note shortly, developed the data Gelbach uses to calculate his lower-bound measure—found an increase of only 0.2% in job discrimination cases, an increase of 8.4% in civil rights cases, FJC SECOND STUDY, *supra* note 7, app. A at 7 tbl.A-1, and an increase of roughly 6% in Total Other cases (based on my recalculation of the FJC data to fit Gelbach's Total Other definition). Note here a critically important point: Gelbach's "negatively affected share" is a measure of the proportion of post-*Twiqbal* plaintiffs affected *among plaintiffs who faced 12(b)(6) motions*, *not* the proportion of plaintiffs affected in all filed cases. To derive the latter measure, one would need to multiply Gelbach's shares by 0.06, given the FJC First Study's finding that roughly 6% of all cases filed in the post-*Twiqbal* period drew a 12(b)(6) motion. *See* FJC FIRST STUDY, *supra* note 7, at 9. Doing so will, for some observers at least, make the stakes of the *Twiqbal* decisions seem far less dramatic, as only roughly 1% of plaintiffs, using Gelbach's selection-adjusted measures, have been adversely affected by the decisions.

80. *See* Gelbach, *Locking*, *supra* note 7, at 2328 n.180, 2329 & tbl.4 (noting and defending the use of estimates from Table A-1 of the FJC Second Study to calculate the percentage of cases in which the movant prevailed); *id.* at 2328 ("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 to dismiss]."); *see also* FJC SECOND STUDY, *supra* note 7, at 3 (defining a prevailing party for purposes of the analysis). For a refresher on why this is important, see *supra* notes 62-64 and accompanying text (advocating coding for full, plaintiff-excluding dismissals). The FJC's grant-rate calculation does, however, take account of motions granted with and without leave to amend and also follows the former to learn whether or not the plaintiffs were able to successfully replead, thus eliminating that

proves to be more than just an academic critique. Plugging the FJC Second Study's estimates of the post-*Twiqbal* 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% among "Total Other" case types.[81]

To be sure, neither of the measurement approaches used to derive these competing estimates is ideal. The grant-as-to-one-or-more-claims approach Gelbach uses 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 peripheral defendants in multidefendant cases.[82] And yet, Gelbach has rightly noted that keying instead on a 12(b)(6) grant's plaintiff-excluding effect would not capture the *Iqbal* case itself, since plaintiff Javad Iqbal was allowed to proceed against the line-level guards also sued.[83] Further research could determine

---

concern. *See supra* notes 59-61 and accompanying text (noting this problem with respect to other *Twiqbal* empirical efforts).

81. Based on my recalculations using the FJC data, the rates at which judges entirely dismissed at least one plaintiff via 12(b)(6) grants in various case types were higher and lower than they were for full grants dismissing all claims challenged. Specifically, plaintiff-excluding grant rates *rose* 7% post-*Twiqbal* in job discrimination cases (as against only a 0.2% rise in partial-or-full grants according to Gelbach's calculations), *declined* by only a fraction of a percent in civil rights cases (as against the 7.8% rise in partial-or-full grants Gelbach finds), and *rose* 4% in "All Other" cases (as against the 1.1% rise in partial-or-full grants Gelbach finds). *See* Gelbach, *Locking*, *supra* note 7, at 2329 tbl.4. For a graphical depiction of how Gelbach's and my estimates diverge, see Figure 1, below. Note here that the mean-comparison and multivariate results from the FJC Second Study for job discrimination, civil rights, and "Total Other" cases do not vary substantially, and so it does not matter which is used to derive the alternate "Engstrom" lower-bound estimates. This also casts at least some doubt on the concern noted previously about the failure of many *Twiqbal* empirical studies to include covariate controls when deriving estimates. *See supra* notes 51-53 and accompanying text. However, as noted previously (and as reported in the second segment of Figure 1 below), the difference between simple mean comparisons and multivariate estimates *does* seem to matter in non-trivial ways in other models. *See infra* Figure 1 (showing that covariate controls reduce the estimated *Twiqbal* effect from 7.2% to 4.3% in the FJC Second Study's estimate of plaintiff-excluding grants among orders granting all claims).

82. Examples of this latter scenario abound in modern litigation. To note just one, medical malpractice plaintiffs often sue the hospital (on a respondeat superior theory of liability) in addition to the treating physicians, but dismissal of the hospital at the 12(b)(6) stage does not preclude discovery or ultimate recovery and, indeed, may only prove relevant when the treating physicians are underinsured and damages exceed the physicians' insurance liability limits. *See*, *e.g.*, Siggers v. Barlow, 906 F.2d 241, 242-43 (6th Cir. 1990) (noting a jury verdict in plaintiff's favor against individual physician after hospital's dismissal from the litigation).

83. *See* Gelbach, *Locking*, *supra* note 7, at 2328. Query, however, how Iqbal and his counsel viewed their chances of recovering from Ashcroft and Mueller as against the prison guards—and, thus, which defendants they saw as central and which peripheral to that litigation.

the incidence of the over- and underinclusiveness of these or other approaches, perhaps permitting analysts to form a clear judgment about which is more informative. Still, the overinclusiveness of Gelbach's measurement approach and the sharp divergence of the competing estimates—especially the substantial reduction of the affected share in civil rights cases—suggest that his study does not quite capture the extent to which *Twiqbal* is "locking the doors to discovery."[84]

The second potential concern with Gelbach's empirical findings is that one of the FJC datasets on which Gelbach relies in deriving negatively affected

84. Note three further points here. First, it bears emphasis that Gelbach's and my estimates do not differ because of something in the way the statistical analysis is performed, as the data sample and the basic calculation method are the same. (The sole exception here is that Gelbach excludes roughly thirty Americans with Disabilities Act cases from the civil rights category, which I am unable to do without access to the FJC data. *See* Gelbach, *Locking*, *supra* note 7, at 2291 n.93, app. A.) Rather, we are measuring different quantities of interest by, in effect, coding the dependent variable differently. Second, because the competing Gelbach and Engstrom estimates are lower-bound measures, it is not quite right to say that one set of estimates is lower or higher than the other. Indeed, a smaller estimate would not imply that the negatively affected share is lower, but rather than that we cannot say it is not lower. Finally, note that comparing negatively affected shares in the above manner raises an interesting question about whether measuring marginal effects—that is, the increase in the probability of a 12(b)(6) grant—is the best way to gauge *Twiqbal*'s effect in the first place. An alternative way to measure the *Twiqbal* effect is to consider the proportion of post-*Twiqbal* complaints subjected to 12(b)(6) grants that would not have been had the pleading regime not changed. To derive this measure, we would divide the marginal change in the post-*Twiqbal* grant rate—again, the negatively affected share—by the post-*Twiqbal* grant rate. *See* Gelbach, *Locking*, *supra* note 7, at 2332-33 (performing a similar calculation). The advantage of this approach is that it works in tandem with, and thus takes account of, differences in the underlying grant rate across case types or grant types (e.g., full or partial grants). This is potentially quite important: as an example, taking Gelbach's estimate of the post-*Twiqbal* "negatively affected" share in job discrimination cases (15.3%, as reported in Figure 1) and noting as well his finding (via the FJC Second Study) that post-*Twiqbal* judges granted 12(b)(6) motions as to one or more claims 61.1% of the time in job discrimination cases, *see id.* at 2329 tbl.4, we would conclude that roughly a quarter (15.3/61.1 = 25.0) of post-*Twiqbal* cases saw 12(b)(6) grants as to one or more claims that would not have occurred had the prior *Conley* regime remained in place. Now compare this to the same calculation for the revised, "Engstrom" estimate of the "negatively affected" share in job discrimination cases (i.e., 10.8%), again as reported in Figure 1. Because the rate at which judges hearing job discrimination cases granted 12(b)(6) motions with plaintiff-excluding effect was far lower than the grant rate dismissing one or more claims (23.0% versus 61.1%, based on the FJC Second Study's results), the proportion of post-*Twiqbal* cases in which plaintiffs suffered entire dismissals who would not have done so had *Conley* remained the operative pleading standard was nearly 50% (10.8/23.0 = 47.0). Viewed in *marginal* terms, Gelbach's estimate appears to be the larger one, since for every hundred 12(b)(6) motions filed, he finds that *Twiqbal* produced at least fifteen more 12(b)(6) grants, whereas the Engstrom estimate shows at least ten extra grants. But calculated as a *proportion* of post-*Twiqbal* grants, one could argue that the Engstrom estimate shows a larger post-*Twiqbal* effect, since nearly *half* of job discrimination plaintiffs who suffered entire dismissals post-*Twiqbal* would not have done so absent a change to the pleading standard, while only *a quarter* of post-*Twiqbal* job discrimination plaintiffs who suffered dismissal as to one or more claims would not have done so absent *Twiqbal*'s elevated pleading standard.

plaintiff shares is flawed in ways that almost certainly inflate his estimates. As noted previously, Gelbach's selection-accounting framework adjusts for selection effects by taking the pre- and post-*Twiqbal* change in 12(b)(6) grant rates and then adding a correction term derived from the pre-*Twiqbal* grant rate and pre- and post-*Twiqbal* 12(b)(6) filing counts.[85] But his measure of post-*Twiqbal* grant rates comes from the FJC's dataset examining orders resolving motions to dismiss between January and June 2010, and a recent investigation by the FJC study's lead author has revealed that a little more than one-quarter of the underlying cases were filed *before* the Court's *Iqbal* decision made clear *Twombly*'s trans-substantive reach.[86] Thus, while Gelbach's framework valiantly adjusts for selection effects, at least some portion of the orders on which his underlying estimates rely are directed at plaintiffs who may have been caught off guard by—and thus filed cases into the teeth of—*Twiqbal*'s elevated pleading standard. As a result, the FJC estimates on which Gelbach relies likely overstate the post-*Twiqbal* change in the 12(b)(6) grant rate, which will in turn inflate Gelbach's own selection-adjusted estimates (and also the alternate "Engstrom" calculations just presented).[87]

In sum, Gelbach's framework is plainly a huge methodological step forward. But, as the Appendix reflects, his empirical estimates may face many of the same shortcomings that afflict *Twiqbal* empiricism more generally: sampling problems, a lack of statistical controls that can neutralize competing explanations for changes in 12(b)(6) grant rates, and an order-level unit of analysis that neglects the critical question of just how many plaintiffs are being barred from the system under *Twiqbal*'s newly constituted pleading regime.

3. *Salutary and non-salutary judicial merits-screening*

A third version of the forest-and-trees problem is the quickest to state but also the most devastating: even a party-level study that focuses on complete (as against partial) dismissals and ultimate (as against initial) 12(b)(6) dispositions and likewise takes account of all possible dynamic litigant responses within the system would *still* not tell us whether judges are using their newfound case-screening powers in ways that increase or decrease social welfare. This is because a simple rise in the motion-to-dismiss grant rate is, as an interpretive

85. For the formal equation, see note 78, above.

86. *See* Cecil, *supra* note 58, at 43 n.160 (observing that 28% of the orders in the dataset meet this criterion).

87. As with the mean-comparison-versus-multivariate issue, the question remains how much this sampling issue matters. Of course, one cannot know for sure how many of these pre-*Iqbal* cases might never have been filed had *Twombly*'s trans-substantive effect been clear, but a back-of-the-envelope calculation using plausible assumptions suggests the problem may inflate Gelbach's lower-bound estimates (which, as noted, range from 15.4% to 21.5%) by two to three percentage points. *See* Gelbach, *Dark Arts*, *supra* note 51, at 13. This is a small but nontrivial change.

matter, equally consistent with socially efficient case screening by trial judges as it is with flawed or biased judicial screening efforts. Put another way, a fully realized empirical estimate of the sort envisioned—but not quite achieved—by Gelbach may tell us *whether* judges are using their newfound dismissal powers, but not *how* they are doing so and to what benefit or cost.[88] Here, then, is the most defeating observation of all, for it suggests that even the best among the Appendix's empirical efforts can offer only limited guidance to a Congress or Advisory Committee considering revising the *Twiqbal* pleading standard.[89]

C. *Does It Matter? A* Twiqbal *Empiricism Meta-Analysis*

Bracketing for now the concern just noted about the judicial will and capacity to perform case screening, what effect do the other measurement, methods, and conceptual concerns canvassed above have in terms of the inferences we can reasonably draw from *Twiqbal* grant-rate studies about the decisions' effects?

88. As noted previously, bottom-line social-welfare judgments are notoriously difficult to make with any precision. *See supra* note 64. Indeed, even a study establishing local judicial capacity to engage in probability screening in resolving motions to dismiss need not imply global efficiency, since the resulting pattern of dismissals could well yield suboptimal levels of deterrence or compensation. For more on this, including the possibility that the *Twiqbal* Court was not instructing lower-court judges to implement a simple probability screen at all, but rather a broader social-welfare judgment about the net benefit or cost of allowing a given case to proceed, see *infra* note 99.

89. Congressional efforts to reconsider and potentially override *Twiqbal* gathered steam in the decisions' immediate aftermath but have since fallen away. *See* Michael R. Huston, Note, *Pleading with Congress to Resist the Urge to Overrule* Twombly *and* Iqbal, 109 MICH. L. REV. 415, 425-27 (2010) (reviewing unsuccessful congressional proposals to overrule the *Twiqbal* decisions).

FIGURE 1
Meta-Analysis of *Twiqbal* Grant-Rate Empiricism

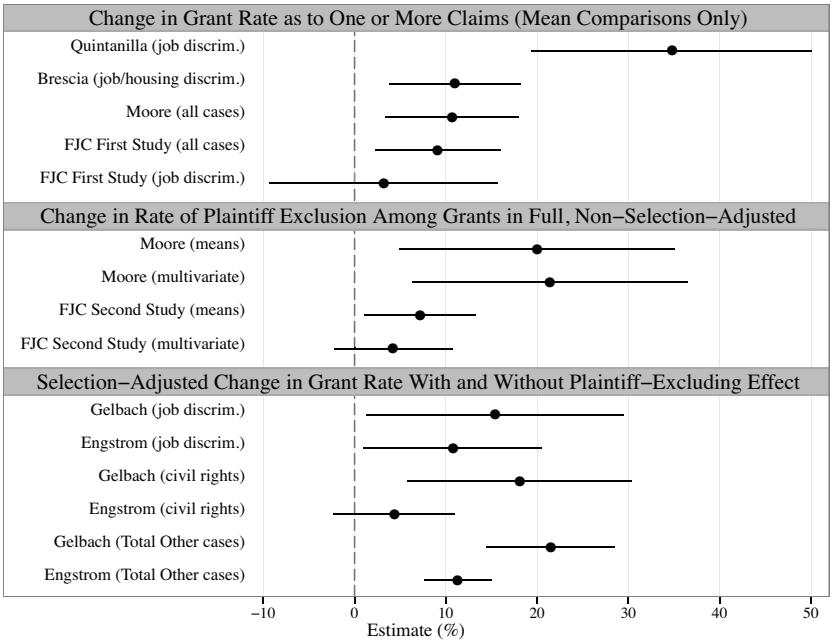

Figure 1 aims to make some progress on that question by offering a meta-analysis of sorts that groups and then arrays results from a number of the Appendix's studies from least to most rigorous in terms of the sampling, measurement, and estimation techniques used and from least to most informative in terms of gauging *Twiqbal*'s impact on plaintiff access to the legal system. More specifically, Figure 1 allocates estimates—including point estimates and 95% confidence intervals[90]—to one of three segments, moving from simple mean comparisons of the pre- and post-*Twiqbal* rate at which orders upon 12(b)(6) motions dismiss one or more claims (the first segment), to estimates of the change in the rate at which orders had plaintiff-excluding effect among fully granted 12(b)(6) motions (the second segment), to selection-adjusted estimates of the change in 12(b)(6) grant rates using Gelbach's grant-as-to-one-or-more-claims approach and the alternate Engstrom approach keyed to a 12(b)(6) grant's plaintiff-excluding effect (the third segment).[91]

90. As is customary with meta-analyses, I report both point estimates and 95% confidence intervals, with the latter meaning we can be 95% confident that the "true" estimate lies somewhere within the line-barred interval.

91. *See supra* notes 81, 84, and accompanying text (describing in detail the differences between these two measurement approaches).

While the resulting analysis can provide only a rough accounting of the full body of *Twiqbal* empiricism,[92] comparing estimates in this way permits two unmistakable conclusions. First, measurement and methods matter, with less rigorous studies analyzing less inclusive data samples tending to report larger *Twiqbal* effects and more rigorous studies analyzing more inclusive data samples tending to report smaller or statistically insignificant effects. The top and bottom studies in the first segment are revealing bookends in this regard, with Victor Quintanilla finding a 35% post-*Twiqbal* increase in 12(b)(6) grant rates as to one or more claims in job discrimination cases using a Westlaw-drawn, citation-keyed sample trimmed to include only orders adjudicating "ambiguous" 12(b)(6) motions challenging a pleading's factual sufficiency, and the FJC First Study finding little or no change in the post-*Twiqbal* rate of such grants using an untrimmed, PACER-drawn near-census of 12(b)(6) orders.[93]

A more focused version of this point is also apparent in the second segment: Hatamyar Moore's less rigorous study (using a Westlaw-based, citation-keyed sample; a case-level analysis that does not track 12(b)(6) grants entered with leave to amend; and a less complete set of covariate controls in the multivariate portion of the analysis) returns estimates of the post-*Twiqbal* change in the rate of 12(b)(6) grants entirely dismissing one or more plaintiffs that are three to five times the estimates returned by the FJC's more rigorous study

92. As with many meta-analyses, myriad differences across the *Twiqbal* grant-rate studies catalogued in the Appendix—including differing case types (all cases, job discrimination, civil rights; exclusion of jurisdictional and fraud cases), plaintiff types (represented versus pro se), and the like—make complete comparability impossible to achieve. This is particularly the case in the first segment, which limits the field to studies reporting estimates for represented (not pro se) plaintiffs in the post-*Iqbal* period (not the *Twombly*-to-*Iqbal* interval, when *Twombly*'s trans-substantive reach was uncertain) to ensure rough comparability, but otherwise mixes and matches grant-rate estimates targeting all cases and particular case types (e.g., job discrimination, or job and housing discrimination combined). This necessarily muddies inferences about the influence of methodological choices on estimates of the *Twiqbal* effect. Note as well that in some instances the estimates presented in Figure 1 have been backed out from the reported results using the number of observations and, where reported, *p*-values or *z*-scores.

93. *See* Quintanilla, *supra* note 7, at 32 (noting the study's aim of "test[ing] the hypothesis that *Iqbal*'s plausibility standard has had a statistically significant effect on Black plaintiffs' claims of race discrimination and racial harassment in ambiguous cases"); *id*. at 33-34 (defining "unambiguous" as including technical dismissals, such as failure to file an EEOC charge as required by statute, or otherwise grounded in "clear rules applied in heuristic fashion"); *supra* note 40 (noting possible omissions from the FJC's intended census of cases). Note that some might consider it unfair to compare Quintanilla's study to the FJC's study in this way, as Quintanilla's study is narrowly focused on the effect of implicit bias and aversive racism on judicial decision-making. *See* Quintanilla, *supra* note 7, at 17-30 (reviewing the social psychology literature on racial bias in forming hypotheses about *Twiqbal*'s effect). Still, comparing the Quintanilla and FJC studies dramatically illustrates the extent to which methods (sampling techniques) and measurement (trimming of the sample to include only certain dismissal rationales) can impact empirical estimates.

(using a PACER-drawn near-census of cases;[94] a party-level unit of analysis that tracks grants entered with leave to amend; and a fuller set of covariate controls in the multivariate analysis).[95] Final confirmation is found in Figure 1's third segment, which reprises the prior analysis showing that replacing the Gelbach measurement approach keyed to grants as to one or more claims with an alternate approach keyed to 12(b)(6) grants with plaintiff-excluding effect (the "Engstrom" approach) yields substantially smaller lower-bound estimates of *Twiqbal*'s effect, particularly among civil rights cases, where the estimate is both small and statistically indistinguishable from zero.[96]

---

94. *But see supra* note 40 (characterizing the FJC dataset as a "near-census" of 12(b)(6) orders in twenty-three district courts based on the possibility that some orders may have been missed during data collection).

95. In order to render the results reported in the multivariate regression models comparable to the mean-comparison findings, I have converted the reported logit coefficients and odds ratios to marginal effects, which are also the more behaviorally interpretable metric. Note here an objection to the way certain *Twiqbal* empirical studies report results. Hatamyar Moore's study repeatedly reports odds ratios but then appears to interpret them in relative risk terms. *See, e.g.*, Hatamyar Moore, *Updated Impact*, *supra* note 7, at 625-26 (reporting an odds ratio of 3.07 and suggesting that this means that a court was "three times more likely" to grant a motion to dismiss). But this is incorrect and misleading, as an odds ratio is, as its name suggests, a ratio of the odds of an event happening to the odds of it not happening. More formally, it is: $[p_2/(1 - p_2)]/[p_1/(1 - p_1)]$. But this is different from *relative risk*, which is merely the ratio of the probabilities of two events happening—for example, the probability that a 12(b)(6) motion will be granted post-*Twiqbal* divided by the probability that the motion will be granted pre-*Twiqbal*, or p_2/p_1. Crucially, and using the example from above, an odds ratio of 3.07 does not necessarily mean that an event is three times more likely to occur. As a concrete example, an increase in the 12(b)(6) grant rate from 50% to 75% would produce an odds ratio of 3.0, since $[0.75/(1 - 0.75)]/[0.5/(1 - 0.5)] = 3$. But in relative risk terms, the event is only 1.5 times more likely to occur, since $0.75/0.5 = 1.5$, and the marginal effect on probability is only 0.25, since $0.75 - 0.5 = 0.25$. Hatamyar Moore's analysis thus risks creating the impression with a less technically adept reader that the post-*Twiqbal* increase in grant rates is much larger than the study in fact finds. *See*, *e.g.*, Suzette M. Malveaux, *The Jury (or More Accurately the Judge) Is Still Out for Civil Rights and Employment Cases Post*-Iqbal, 57 N.Y.L. SCH. L. REV. 719, 742-43 (2013) (reprising Moore's findings in misleading, relative risk terms).

As a final note, and as with the first segment's analysis, the limitations of Figure 1's meta-analysis once more bear emphasis: it is impossible to determine the extent to which the divergence between Hatamyar Moore's results and the FJC Second Study's results are attributable to the measurement and methods and differences noted above or other differences between the studies. For instance, the differences might be explained by the fact that Hatamyar Moore's dependent variable is whether the entire case was dismissed, while the FJC report considers whether one or more plaintiffs was entirely dismissed (though one might expect this to narrow, not widen, the difference between the estimates). For full-scale analyses of a number of other differences between the two studies beyond these basic points, but with few firm conclusions about the likely magnitude or direction of the likely bias resulting from myriad minor study differences, see Hatamyar Moore, *Updated Impact*, *supra* note 7, at 634-51; and Cecil, *supra* note 58, at 21-34.

96. *See supra* notes 81-84 and accompanying text. The standard errors used to generate confidence intervals for the "Engstrom" lower bound estimates in Figure 1's third segment were derived using the complex procedure Gelbach sets forth in an online appendix to

TABLE 1
Proportion of *Twiqbal* Effect Due to Litigant Selection

Study (Case Type)	Judicial Grant Rate Term	Selection Correction Term	Overall *Twiqbal* Effect Estimate	Proportion of Effect Due to Selection
Gelbach (job discrim.)	0.2	15.2	15.4	98.7%
Gelbach (civil rights)	7.8	10.3	18.1	56.9%
Gelbach ("Total Other")	1.1	20.4	21.5	94.9%
Engstrom (job discrim.)	6.7	4.1	10.8	38.0%
Engstrom (civil rights)	0	4.4	4.4	100%
Engstrom ("Total Other")	3.9	7.4	11.3	65.5%

Second, the more rigorous studies in Figure 1's first and second segments suggest that *Twiqbal* has had, at most, a single-digit impact on the observed rate at which judges have granted 12(b)(6) motions in cases where motions to dismiss were filed (i.e., the "judicial behavior" effect), and this is true of orders with and without plaintiff-excluding effect. Yet Table 1 captures a further, and critically important conclusion that follows from a combination of these estimates of *Twiqbal*'s effect on the observed judicial grant rate and the third segment's overall, selection-adjusted estimates of the *Twiqbal* effect: in gauging *Twiqbal*'s impact, litigant selection matters, too, perhaps even more than the observed change in judicial grant rates. Indeed, even deflating Gelbach's selection adjustments slightly to account for the previously mentioned problems with the FJC's motions-filing data,[97] party selection accounts for at least half of the third segment's estimates of the *Twiqbal* effect. Put another way, selection and settlement effects, *not* the more directly observable change in the *judicial* grant rate, may be the more important dynamic in measuring *Twiqbal*'s effect on plaintiff access to the legal system.

III. IS THE BLOOM OFF THE ROSE? LESSONS FOR EMPIRICAL STUDY OF CIVIL PROCEDURE

Part II's critique exposes significant problems with existing *Twiqbal* empirical efforts. Clearly, much work remains to be done if we are to move closer to solving the *Twiqbal* puzzle. And yet, the above analysis is not without bright spots in terms of how to go about it. For instance, Gelbach's work on litigant selection provides an accessible and fully implementable framework that future empirical researchers analyzing procedural change, whether in the *Twiqbal* context or beyond, ignore at their peril. Similarly, anatomizing *Twiqbal*

his *Locking* article. *See* Gelbach, *Locking*, *supra* note 7, app. B, *available at* http://www.yalelawjournal.org/images/documents/gelbach_appendix_b.pdf.

97. *See supra* notes 85-87 and accompanying text.

grant-rate studies helps us to see how the two studies reviewed back in Part I—including Gelbach's more recent and preliminary study using summary judgment grant rates to measure judicial merits-screening capacity at the motion-to-dismiss stage and also the Boyd et al. effort to map the post-*Twiqbal* pleading landscape[98]—constitute a welcome rechanneling of scholarly effort away from isolating a judicial behavior response in *Twiqbal*'s wake and toward a more expansive consideration of the decisions' systemic and social welfare effects.[99]

But if canvassing the best and worst of *Twiqbal* empiricism points the way to more productive approaches to empirical analysis of the *Twiqbal* puzzle, then it also raises some deeper, and at times disquieting, questions. This Part steps back from Part II's fine-grained critique of *Twiqbal* empiricism and of-

98. *See supra* Part I.A.

99. This is not to suggest that the two studies are immune from criticism. While the embryonic nature of Gelbach's summary judgment study, *see* Gelbach, *Material Facts*, *supra* note 7, makes a full-scale critique inappropriate, at least two potential concerns stand out. First, Gelbach's model is limited by its assumption that judges implementing *Twiqbal*'s plausibility standard will apply a simple probability screen keyed to the likelihood that the plaintiff will ultimately discover inculpating evidence and make out her claim. But the *Twiqbal* Court specifically disclaimed that it was directing trial courts to do any such thing. *See* Ashcroft v. Iqbal, 556 U.S. 662, 678-79 (2009); Bell Atl. Corp. v. Twombly, 550 U.S. 544, 556 (2007). Rather, the Court's plausibility concept, some recent commentators have suggested, admits of both probability *and* consequences, requiring trial judges to apply a broad balancing test that includes both the likelihood that discovery will reveal inculpating facts and also the likely litigation and other costs that will be incurred in getting there. *See, e.g.*, Louis Kaplow, *Multistage Adjudication*, 126 HARV. L. REV. 1179, 1256-57 (2013). The problem here is that, if post-*Twiqbal* trial judges are screening cases based on something resembling their projected social value (rather than their simple probability of success), then we lose any firm prediction about the direction of the post-*Twiqbal* shift in summary judgment grant rates. To that extent, Gelbach's innovative identification strategy can offer evidence on only the most simplistic account of what trial judges do—or have been instructed to do—under *Twiqbal*.

A second potential problem is that Gelbach's study does not in its current form take account of the dynamic litigant responses in *Twiqbal*'s shadow that his *Locking* study so elegantly addresses. In particular, Gelbach's new study appears to rely on an identifying assumption that a complaint's survival of a motion to dismiss under *Twiqbal*'s heightened pleading standard will not alter the parties' settlement calculus prior to summary judgment relative to a *Conley* notice-pleading world. Put more formally, his analysis assumes that a judge's decision at the motion-to-dismiss stage and her later decision upon a motion for summary judgment are not correlated—and, just as crucially, that litigants will not see them as linked, either. But this assumption is unlikely to hold in any real-world litigation context. In a *Twiqbal* world, a judge's 12(b)(6) denial may well reveal valuable judge-specific information about what precise legal standard will apply and what factual showing will be necessary to meet it, allowing plaintiffs and defendants alike to update (and refine) their prior assessment of their likelihood of prevailing, potentially altering litigation and settlement behavior. *See, e.g.*, Christina L. Boyd & David A. Hoffman, *Litigating Toward Settlement*, J.L. ECON. & ORG. (forthcoming 2013) (manuscript at 26), *available at* http://ssrn.com/abstract=1649643 (noting that dispositive motions practice can "unlock[] information that the parties and the court otherwise would not share with each other"); *see also* George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1, 4, 12-24 (1984) (modeling settlement as a function of litigant expectations).

fers, albeit briefly, some concluding reflections on what that exercise might tell us, about both the health of the ELS movement at its current stage of development and the shape of empirical study of civil procedure going forward.

A. *The Double-Edged Sword of Democratization*

Perhaps the most obvious questions arising out of Part II's critique of *Twiqbal* empiricism concern the effects of the broadening ranks of ELS practitioners. The potential costs of that trend should by now be coming into focus. One is the steep opportunity costs that arise from the unproductive diversion of scholarly capacity into time-consuming and resource-intensive empirical projects. It is hard to imagine a better illustration of this than the tens of thousands of hours of research effort reflected in many of the Appendix's catalog of *Twiqbal* studies.

Low-grade empirical research may also be counterproductive in a more direct sense. An important part of ELS's promise at the dawn of the movement was that empirical research—even relatively simple descriptive work—could discipline public debate over litigation by deterring the more overheated claims made by the Chamber of Commerce, the American Tort Reform Association, or the plaintiffs' bar.[100] The problem is that the wildly divergent empirical findings—like those catalogued in the Appendix and graphically illustrated in Figure 1—may have little disciplining effect. Indeed, such efforts may achieve just the opposite, muddying debate and liberating public actors from any data-based accountability at all.[101]

A final potential cost is just as serious—and also somewhat unique to empirical study of civil procedure and civil litigation. Specifically, poorly executed empirical studies that purport to measure the effects of procedural or other legal change within litigation regimes may well *shape* litigant and judge perceptions as much as (or even more than) they *reveal* them.[102] As a concrete example drawn from the above discussion, low-quality empirical work risks exacerbating selection effects and, in particular, the chilling effect on the claiming behavior of aggrieved parties that some worry will be *Twiqbal*'s most significant and enduring effect.[103]

---

100. *See* Eisenberg, *supra* note 10, at 1736 (arguing that even basic empirical data can expose "the shoddy empirical claims" made by politically motivated actors such as the Chamber of Commerce and the American Tort Reform Association).

101. One reason this is so is that, as Lee Epstein and Gary King have put it, the "staying power of flawed and discredited legal studies can be extraordinary." Lee Epstein & Gary King, *The Rules of Inference*, 69 U. CHI. L. REV. 1, 17 n.42 (2002).

102. *Cf. id.* at 7-9, 12 n.29 (asserting that legal scholarship, particularly empirical work, has more power to shape behavior and influence public policy relative to other academic disciplines because of judge and litigant reliance on it).

103. *See supra* note 74 and accompanying text (noting commentators' concern about *Twiqbal*'s chilling effect on claiming behavior). The explanation advanced by one of the

These costs can be substantial, and an anodyne response might be to return empirical inquiry to the hands of more sophisticated technicians. But even if such an option were available, it would be a mistake to conclude that more technocracy is the cure. As an initial matter, the Appendix's canvass of *Twiqbal* empirical efforts suggests that one of the most significant obstacles to higher-quality empirical research within the civil procedure space remains data availability and the sampling problems that flow therefrom. Most frustrating, and as Part I noted, it is the judiciary itself—and, more specifically, chief district judges—who have maintained the principal barrier to more methodologically sound empirical legal research on the workings of the civil justice system by refusing to use their statutory discretion to grant academic fee waivers for research conducted using the PACER system.[104]

Nor, with the benefit of some broader perspective, is it clear that what most ails *Twiqbal* empiricism is a failure to utilize more sophisticated statistical methods. In reality, ELS has, as an intellectual movement, been on something of a collision course in recent years. Indeed, electronic docketing has lowered barriers to entry for legal scholars conducting empirical research just as the movement's most sophisticated practitioners have, paralleling a broader move in the social sciences, sought to effect a "credibility revolution" in how such research is performed.[105]

Appendix's empirical authors is particularly revealing in this regard. After conceding concern about use of a Westlaw-derived sample, Quintanilla offers the following rationalization:

> The study, therefore, does not seek to establish the absolute rate of dismissals in all decisions, or to measure the absolute number of Rule 12(b)(6) motions decided before and after *Iqbal*. The study remains significant, however, because jurists and advocates do not form impressions about what the law is from inaccessible law; if a disparate effect is demonstrated in available law, that effect will have practical significance for how jurists and advocates handle cases.

Quintanilla, *supra* note 7, at 31 n.209. The point here appears to be that the author's reporting of statistics based on published-but-concededly-unrepresentative dispositions is policy-relevant because judges and litigants can see only the visible "tip" of published decisions and will take decisional cues from them. But if the goal is thus to predict litigant and judicial behavior by isolating decisional biases, then one wonders why a survey of judges or practitioners or experimental research would not be the better course. *See, e.g.*, THOMAS E. WILLGING & EMERY G. LEE III, FED. JUDICIAL CTR., IN THEIR WORDS: ATTORNEY VIEWS ABOUT COSTS AND PROCEDURES AND FEDERAL CIVIL LITIGATION 25 (2010) (surveying plaintiff- and defense-side attorneys and finding few reports of any *Twiqbal* effect). And if the goal is to correct decisional biases by showing the divergence between the "true" state of the world and the apparent (but erroneous) *Twiqbal* effect as measured via "available" law, then the study falls short, as it does nothing to compare its findings to the "true" state of the world. In so doing, the study risks contributing to the same decisional bias it seeks to expose, thus exacerbating litigant selection effects.

104. *See supra* note 30.

105. *See generally* Joshua D. Angrist & Jörn-Steffen Pischke, *The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics*, J. ECON. PERSP., Spring 2010, at 3 (2010) (sketching tenets of the "credibility revolution" in empirical methods). On the role of electronic docketing in expanding the accessibility of empirical legal research, see *supra* notes 20-24 and accompanying text.

But notice two critical tenets of that revolution. First, a growing consensus holds that valid statistical inference can best be achieved not via acrobatic math after data has been collected and coded but rather careful research design before data collection begins.[106] It follows that many legal scholars' lack of formal methods training, at least of the pure econometric variety, need not present an insuperable barrier to the continued expansion of the field.

A second, and related, tenet is that the increasing technical sophistication of empirical legal research presents risks as well as benefits. Chief among the concerns is that a move toward use of computer-automated systems to create ever-larger datasets will crowd out qualitative institutional insight—and, more specifically, lawyerly understanding and judgment—in the formation of hypotheses, the construction of data samples, and the coding of variables.[107] The result may be a "naive" empiricism that is prone to basic interpretive errors and no more likely to generate valid inferences about the complex interactions of procedure and substantive justice than qualitative surveys of doctrinal developments or practitioners' ground-level, gestalt sense of things.[108]

An illuminating example as to the latter point is Dodson's *Twiqbal* grant-rate study, which, though afflicted by some of the methods and measurement concerns noted above, also displays far greater lawyerly engagement with the cases in the study sample than most or all of the Appendix's other studies.[109] The result is a rich descriptive portrait of the *Twiqbal* transition and a set of insights that are unique among *Twiqbal* empirical efforts, including fascinating findings on what were previously termed "decisional hydraulics"—that is, the possibility that post-*Twiqbal* trial judges are dismissing claims on factual sufficiency grounds that they previously reached to dismiss on legal sufficiency

106. *See* Daniel E. Ho & Donald B. Rubin, *Credible Causal Inference for Empirical Legal Studies*, 7 ANN. REV. L. & SOC. SCI. 17, 17 (2011) ("Research design trumps methods of analysis.").

107. *See, e.g.*, Eisenberg, *supra* note 10, at 1729 ("[S]cholars with limited legal training sometimes misdescribe the system, as illustrated by inflated claims about the settlement rate for filed cases, or get the law wrong." (footnote omitted)).

108. For an early statement of the perils of "naive empiricism," see Willard Hurst, *Perspectives upon Research into Legal Order*, 1961 WIS. L. REV. 356, 365.

109. Another way of putting this point is that Dodson's study, more than the others, engages in actual "content analysis" of the cases by considering the court's substantive legal reasoning rather than merely recording outcomes that can be discerned from docket sheets alone. *See* Hall & Wright, *supra* note 22, at 72-73 (contrasting "docket analysis" with "content analysis," with the former coding "only for information about cases—such as subject matter, parties, and basic outcomes—that could be obtained from docket sheets or brief abstracts," as against the latter's effort to reach "the substance of judicial reasoning as expressed through the legal and factual content of written opinions"). For another example of a *Twiqbal* study that supplements quantitative analysis of grant rates with useful content analysis, see Brescia, Iqbal *Effect*, *supra* note 7, at 279-80. That said, some question the capacity of even the more searching content analyses to generate useful inferential judgments about judicial behavior. *See* Howard Gillman, *What's Law Got to Do With It? Judicial Behavioralists Test the "Legal Model" of Judicial Decision Making*, 26 LAW & SOC. INQUIRY 465 (2001) (book review).

grounds.[110] A legal empiricist with a lesser command of civil procedure and litigation practice than Dodson could not produce such insights or articulate them as effectively.

Finally, any analysis of the pros and cons of ELS's deepening penetration into the ranks of legal scholars must take account of the complex nexus of empirical legal research and the political system.[111] Historically, empirical research on civil procedure has issued from one of two quarters. The first was transparently partisan efforts to advocate for rule reforms on behalf of specific legal or client constituencies.[112] The second source was a relatively limited set of large-scale research initiatives undertaken with substantial (and often public) funding.[113] Only rarely has significant empirical research, at least in the era before electronic docketing, come about through the mania of individual researchers.[114] The promise, then, of the broadening of ELS practitioner ranks on display in the recent spate of *Twiqbal* studies is a more robust form of this third source of empirical research and, with it, a body of empiricism that is fully decoupled from any organizational agenda.

---

110. *See supra* note 49 and accompanying text.

111. *See* Bryant G. Garth, *Observations on an Uncomfortable Relationship: Civil Procedure and Empirical Research*, 49 ALA. L. REV. 103, 113-17 (1997) (noting the complex and often tense relationship between empirical research on civil procedure and political actors).

112. For a powerful indictment along these lines, see Carrie J. Menkel-Meadow & Bryant G. Garth, *Civil Procedure and Courts*, *in* THE OXFORD HANDBOOK OF EMPIRICAL LEGAL RESEARCH, *supra* note 12, at 679, 681-90 (offering a history of research in civil process and procedure throughout the twentieth century and noting the "systematic structural tilt toward political uses of that research"); *see also* Eisenberg, *supra* note 10, at 1736 ("[T]he shortfall in reliable information about the legal system allows self-interested parties to fill the information gap with biased studies marketed as neutral social science.").

113. For examples of the large-scale, heavily funded research initiatives that dominated the field throughout the twentieth century, see JAMES S. KAKALIK ET AL., RAND INST. FOR CIVIL JUSTICE, JUST, SPEEDY, AND INEXPENSIVE? AN EVALUATION OF JUDICIAL CASE MANAGEMENT UNDER THE CIVIL JUSTICE REFORM ACT (1996) (summarizing the results of a $4.5 million research venture examining the implementation of the Civil Justice Reform Act's requirement that each federal district court develop a case management plan to reduce litigation costs and delay); MAURICE ROSENBERG, THE PRETRIAL CONFERENCE AND EFFECTIVE JUSTICE: A CONTROLLED TEST IN PERSONAL INJURY LITIGATION (1964) (reporting results of a foundation-funded study of litigation case management); DAVID M. TRUBEK ET AL., CIVIL LITIGATION RESEARCH PROJECT: FINAL REPORT (1983) (presenting empirical results from the Civil Litigation Research Project, a joint venture between the University of Wisconsin and the University of Southern California, as funded by the Office for Improvements in the Administration of Justice, U.S. Department of Justice).

114. For representative examples of influential individual-researcher-driven projects beginning with early Legal Realist studies of judicial administration, see CHARLES CLARK & HARRY SHULMAN, A STUDY OF LAW ADMINISTRATION IN CONNECTICUT: A REPORT OF INVESTIGATION OF THE ARTICLES OF CERTAIN TRIAL COURTS OF THE STATE 1 (1937) (reporting results of a large-scale study of court administration in Connecticut state courts between 1919 and 1932, which began as an individual research effort and only later secured funding from a foundation); William O. Douglas & J. Howard Marshall, *A Factual Study of Bankruptcy Administration and Some Suggestions*, 32 COLUM. L. REV. 25 (1932).

B. *The Way Forward*

All of this leads to a final question: what should empirical study of civil procedure look like going forward? A full answer to that sprawling question is clearly beyond the scope of the present inquiry. Readers interested in developing a more encompassing sense of where civil procedure empiricism has been and where it might go should instead consult several excellent contributions to *The Oxford Handbook of Empirical Legal Research*.[115]

Still, the analysis thus far suggests some broad prescriptions that legal scholars interested in conducting empirical research in the civil procedure space should, if they have not already done so, take to heart. Some of these will not surprise—and require little elaboration following Part II's full-dress critique of *Twiqbal* grant-rate studies. One is greater methodological rigor, particularly as to data collection. Indeed, perhaps the greatest marginal improvement in civil procedure empiricism going forward will come with fully random samples that can inoculate empirical findings from sampling bias concerns. Similarly, civil procedure empiricism, as with any empirical research area, would plainly benefit from a better alignment of research questions and research design. If the goal is to measure the effect of procedural changes on plaintiff access to the legal system, then empiricists should collect, code, and interrogate data with that end in mind. If the goal, by contrast, is to measure the judicial behavior effect of those same procedural changes—perhaps as a means of testing hierarchical relations between upper and lower courts[116]—then this will likely require an entirely different research approach. This starts with framing the research question as precisely as possible.

But beyond these perennial criticisms, this Essay's deep dive into *Twiqbal* empiricism suggests some further and more civil-procedure-specific prescriptions that have not drawn nearly as much attention in the growing methodological metaliterature on empirical legal studies.[117] First, civil procedure

---

115. *See, e.g.*, Sharyn Roach Anleu & Kathy Mack, *Trial Courts and Adjudication*, *in* THE OXFORD HANDBOOK OF EMPIRICAL LEGAL RESEARCH, *supra* note 12, at 545; Herbert M. Kritzer, *The (Nearly) Forgotten Early Empirical Legal Research*, *in* THE OXFORD HANDBOOK OF EMPIRICAL LEGAL RESEARCH, *supra* note 12, at 875; Menkel-Meadow & Garth, *supra* note 112. Another useful source setting forth the large body of civil procedure studies performed by the Federal Judicial Center in recent decades, many of which sometimes go overlooked in literature reviews, is Thomas E. Willging, *Past and Potential Uses of Empirical Research in Civil Rulemaking*, 77 NOTRE DAME L. REV. 1121, 1147-48 (2002).

116. *See supra* note 54.

117. For an initial call for "greater self-conscious attention to methodology in legal studies," including scholarship devoted to purely methodological issues in conducting empirical legal research, see Epstein & King, *supra* note 101, at 6-7, 11. For leading examples of studies heeding that call with respect to measurement of judicial ideology and decisionmaking, see Joshua B. Fischman & David S. Law, *What Is Judicial Ideology, and How Should We Measure It?*, 29 WASH. U. J.L. & POL'Y 133 (2009); Daniel E. Ho & Kevin M. Quinn, *How Not to Lie with Judicial Votes: Misconceptions, Measurement, and Models*, 98 CALIF. L. REV. 813 (2010).

empiricism would profit from more careful and user-friendly presentation of findings. In particular, more accessible explanations of substantive results—for instance, reporting findings as marginal effects rather than odds ratios[118]—will avoid needless confusion and are especially important in the civil procedure space in light of the likely greater sensitivity of litigants, whether parties or counsel, to empirical findings compared to primary actors in other legal and policy areas.[119]

Second, in drawing research questions and design into better alignment, empirical legal scholars in the civil procedure space should consider a wider menu of approaches and techniques. This may seem like a throwaway point, as empirical study of civil procedure has always been, and will continue to be, pluralistic rather than monolithic. As before, empirical work going forward will no doubt include plenty of *intra*system studies of the effect of rule choices on systemic design values (efficiency, accuracy, fairness, access, decisional legitimacy) akin to Part II's *Twiqbal* empirical efforts, and also some *inter*system (whether cross-state or cross-national) studies that do the same. And it will surely include, continuing a wider trend in the social sciences, substantial experimental research—including "laboratory" simulations and perhaps even controlled field experiments—to better understand likely litigant responses to different rule regimes.[120]

But assessing *Twiqbal* empiricism brings to mind other, smaller-scale approaches that remain untapped in civil procedure empiricism despite their potential value and ease of implementation. In particular, it is striking that, though many *Twiqbal* grant-rate study authors purport to seek to measure a judicial behavior response to the decisions, none deploys the methodological approach that seems best adapted to that task: using "matching" techniques to prune data prior to statistical estimation so that the pairs of cases that remain are as similar

---

118. *See supra* note 95 (noting interpretive confusion where coefficients are reported as odds ratios); *see also* Hubbard, *supra* note 7, at 54-56 tbls.4-6 (reporting logit coefficients as marginal effects).

119. *See supra* notes 102-103.

120. *See, e.g.*, Laurens Walker, *Perfecting Federal Civil Rules: A Proposal for Restricted Field Experiments*, 51 LAW & CONTEMP. PROBS. 67, 84-85 (1988) (calling for field experiments as a way to make progress on understanding the effect of rule changes). As an example, any effort to break the current theoretical impasse on the effect of fee-shifting and offer-of-judgment rules will likely involve experimental simulations in addition to quasi-experimental observational studies. *See* Avery Wiener Katz & Chris William Sanchirico, *Fee Shifting in Litigation: Survey and Assessment* 2 (Inst. for Law & Econ., Working Paper No. 10-30, 2010), *available at* http://ssrn.com/abstract=1714089 ("[T]he current state of economic knowledge does not enable us reliably to predict whether a move to fuller indemnification would raise or lower the total costs of litigation, let alone whether it would better align those costs with any social benefits they might generate."); *id.* at 34 (noting the "relative lack of systematic empirical investigation" of questions relating to fee shifting); *see also* Herbert M. Kritzer, *Lawyer Fees and Lawyer Behavior in Litigation: What Does the Empirical Literature Really Say?*, 80 TEX. L. REV. 1943, 1948 (2002) (noting "surprisingly little agreement" among scholars about the effects of different fee-shifting regimes).

as possible across treatment and control groups.[121] More concretely, an analyst could draw a random sample of pre- and post-*Twiqbal* cases asserting a particular type of Title VII claim and then match like cases in the two samples on the basis of as many relevant, micro-level case attributes as can be gleaned from docket materials prior to estimating the post-*Twiqbal* change in grant rates. To be sure, such an approach would be vulnerable to criticism regarding the accuracy of matching like cases from docket materials alone.[122] But the inferences one could draw about judicial implementation of *Twiqbal* from a well-executed version of such a study would surely be no less valid, and likely far more valid, than the existing raft of studies reporting a non-selection-adjusted post-*Twiqbal* change in 12(b)(6) grant rates.[123] Just as important, note that a matching approach of this sort would rely, quite heavily, on lawyerly expertise and judgment, thus capitalizing on the comparative advantages of *lawyer* empiricists as against those with other kinds of disciplinary training—and might well produce Dodson-like insights in the process.

A final prescription returns us to a point raised much earlier in connection with Part I's review of the Gelbach and Boyd et al. studies: whatever precise forms empirical study of civil procedure takes going forward, it will ideally include at least two components. The first is narrowly targeted, Gelbachian research efforts that test discrete hypotheses about the effects of different procedural rules. But one hopes the field will not thereby ignore the "mapping" studies of the Boyd et al. sort. This is no idle concern. No large-scale empirical study of the American civil justice system has been undertaken since at least the RAND study of the Civil Justice Reform Act in the 1990s.[124] And

121. This approach to estimating the effect of a treatment, policy, or other intervention is commonly referred to as "propensity score matching" and entails creating a sample of units that received the treatment that is as close as possible to the sample of units that did not receive the treatment along all observed covariates as a way to mimic randomization. *See* Daniel E. Ho et al., *Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference*, 15 POL. ANALYSIS 199 (2007) (offering a sophisticated but accessible how-to guide to this approach).

122. Matching of this sort would plainly be an inexact form of matching, as it would depend on the coder's judgment about the meritoriousness—and thus 12(b)(6) survivability—of a case based on information gleaned from docket materials. For related discussion of how to apply matching techniques to real-world data where exact matches among cases cannot be found, see Stefano M. Iacus et al., *Causal Inference Without Balance Checking: Coarsened Exact Matching*, 20 POL. ANALYSIS 1 (2012).

123. The clear tradeoff of a matching approach is the necessary sacrifice of out-of-sample validity for in-sample validity. In less technical terms, the benefit of such an approach is quasi-experimental control: the only features that would differ between the two groups of cases are the changed standard and the passage of time, thus isolating the judicial behavior response. The cost is that any systematic changes in characteristics of the pool of cases post-*Twiqbal* would be ignored in the analysis, precluding inferences about how much the observed judicial behavioral response mattered in terms of real-world litigant fortunes given selection and settlement effects and longer-term litigation trends.

124. *See supra* note 113 (listing past large-scale civil justice studies, including RAND's evaluation of the Civil Justice Reform Act of 1990). The same is true of so-called "civil

surprisingly few researchers have attempted synoptic accountings of the civil litigation system even since the advent of electronic docketing lowered the barriers to doing so.[125]

Part of the reason here may be increasingly scarce public funding in an era of fiscal austerity, or the fact that the litigation wars have receded from their 1990s peak, reducing the direct political salience of such projects.[126] Perhaps as well the steady stream of mostly descriptive FJC reports in specific procedural or litigation areas, often at the behest of the Advisory Committee,[127] has reduced the perceived returns to larger-scale, resource-intensive projects. Yet mapping exercises are every bit as critical to understanding rule choices as more targeted studies of the Gelbach sort. As a number of commentators have noted, the true stakes of procedural choices will often turn on who is using the courts in the first place (business organizations, individuals) and toward what ends (as an extension of business strategy, to vindicate constitutional rights).[128]

---

needs" studies. *See* AM. BAR ASS'N, LEGAL NEEDS AND CIVIL JUSTICE—A SURVEY OF AMERICANS: MAJOR FINDINGS FROM THE COMPREHENSIVE LEGAL NEEDS STUDY (1994); BARBARA A. CURRAN, THE LEGAL NEEDS OF THE PUBLIC 137 (1977). *See generally* Rebecca Sandefur, *Money Isn't Everything: Understanding Moderal Income Households' Use of Lawyers' Services*, *in* MIDDLE INCOME ACCESS TO JUSTICE 222, 224 (Michael Trebilcock et al. eds., 2012) ("The last truly comprehensive surveys of public experience with civil justice problems are more than three decades out of date, conducted in the 1970s.").

125. Relatively rare exceptions include Hoffman, *supra* note 20 (performing a "docketology" study of the factors that affect opinion writing); and Gillian K. Hadfield, *Exploring Economic and Democratic Theories of Civil Litigation: Differences Between Individual and Organizational Litigants in the Disposition of Federal Civil Cases*, 57 STAN. L. REV. 1275 (2005) (mapping trends in litigant identity, particularly individual and organizational litigants, within the federal civil litigation system).

126. For excellent accounts of political battles over litigation during the 1980s and 1990s in particular, see THOMAS F. BURKE, LAWYERS, LAWSUITS, AND LEGAL RIGHTS: THE BATTLE OVER LITIGATION IN AMERICAN SOCIETY (2002); Sean Farhang, *Litigation and Reform*, *in* THE POLITICS OF MAJOR POLICY REFORM IN POSTWAR AMERICA (Jeffrey A. Jenkins & Sidney M. Milkis eds., forthcoming 2013), *available at* http://ssrn.com/abstract=2184562.

127. *See FJC Studies and Related Publications*, U.S. CTS., http://www.uscourts.gov/RulesAndPolicies/rules/archives/fjc-studies-and-related-publications.aspx (last visited June 9, 2013).

128. As Gillian Hadfield eloquently puts it:

> The issues at stake in our understanding of what is happening to civil cases and the efforts to craft alternatives to traditional civil litigation . . . absolutely require that we differentiate between litigants, between legal functions, and between the different goals of our legal system. It may be that the disappearance of public civil trials to resolve commercial contract disputes is of no consequence; indeed, it may be an efficient response to the increasing cost of the public system. The same cannot be said of the disappearance—if it is a real phenomenon—of public adjudication of civil rights or the claims of individuals about the misconduct of public or corporate actors.

Hadfield, *supra* note 125, at 1280; *see also* Menkel-Meadow & Garth, *supra* note 112, at 698-99 (reviewing evidence and concluding that "U.S. federal courts may be playing a very different role today than they were a generation ago, refereeing complex business disputes and managing routine matters, rather than enunciating great constitutional principles"). *See generally* Richard Abel, *Forecasting Civil Litigation*, 58 DEPAUL L. REV. 425 (2009) (delineating factors that shape civil litigation flows).

To repeat Part I's framing, only through methodological cross-pollination—a dynamic working back and forth between more targeted hypothesis tests and more expansive mapping projects—can we achieve a true blossoming of ELS in the civil procedure space.

APPENDIX
Empirical Studies of *Twiqbal*'s Effect on 12(b)(6) Grant Rates at a Glance

Study Name	Unit of Analysis	Case Type Limitations	Random Sample or Case Census?	Multiple Regression with Covariate Controls?	Isolates Plaintiff-Excluding Dismissals Without Leave to Amend?	Tracks Dis-missals with Leave to Amend?	Accounts for Litigant Selection / Settlement Effects?	Core Findings
Hannon (2008)	Order	Excludes fraud, *in forma pauperis*, pro se cases	Neither (Westlaw search keyed to order's citation of *Conley* or *Twombly*)	Yes, but no modeling details provided	No (codes orders granting in full, granting in part, or denying motions, but does not gauge plaintiff exclusion or distinguish between grants with and without leave to amend)	No	No	Post-*Twombly*/pre-*Iqbal* increase in 12(b)(6) grant-in-full rates in civil rights cases of 11% (mean comparisons only), but no material difference among non-civil-rights cases.
Seiner (2009)	Order	Title VII cases only	Neither (Westlaw search keyed to order's citation of *Conley* or *Twombly*)	No	No (codes orders granting in full, granting in part, or denying motions, but does not gauge plaintiff exclusion or distinguish between grants with and without leave to amend)	No	No	Post-*Twombly*/pre-*Iqbal* increase in 12(b)(6) grant rates from 55% to 57% (in full) and 75% to 78% (in full or in part) (mean comparisons), but not statistically significant using Fisher's exact test.
Seiner (2010)	Order	ADA cases only	Neither (Westlaw search keyed to order's citation of *Conley* or *Twombly*)	No	No (codes orders granting in full, granting in part, or denying motions, but does not gauge plaintiff exclusion or distinguish between grants with and without leave to amend)	No	No	Post-*Twombly*/pre-*Iqbal* increase in 12(b)(6) grant rates from 54% to 65% (in full) and 64% to 79% (in full or in part) (mean comparisons), but not statistically significant using Fisher's exact test.

Study Name	Unit of Analysis	Case Type Limitations	Random Sample or Case Census?	Multiple Regression with Covariate Controls?	Isolates Plaintiff-Excluding Dismissals Without Leave to Amend?	Tracks Dis-missals with Leave to Amend?	Accounts for Litigant Selection / Settlement Effects?	Core Findings
Hatamyar (2010)	Order / Case	Excludes *sua sponte* reviews of prisoner suits, jdx dismissals, fraud cases	Neither (Westlaw search keyed to order's citation of *Conley* or *Twombly* / *Iqbal*)	Yes (controls for circuit, case type, judge type, pro se, class action)	Yes (separately codes whether case was "entirely dismissed" by 12(b)(6) order without leave to amend)	No	No	Post-*Twombly* increase in grant rates from 74% to 79% (in full or in part, mean comparisons) and also in grants in full with leave to amend (multivariate, but no marginal effects reported), but no other statistically significant differences in 12(b)(6) grant rates, including the proportion of grants in full "entirely dismiss[ing]" case.
Quintanilla (2011)	Order	Race-based Title VII and § 1981 cases only, excluding "non-ambiguous" claims (e.g., technical failures to satisfy Title VII prerequisites) and retaliation claims	Neither (unspecified "[b]road Westlaw searches" only)	No (logistic regression results reported only for analysis of impact of judge race and plaintiff gender on 12(b)(6) grant rates)	No (codes orders granting in full, granting in part, or denying motions, but does not gauge plaintiff exclusion or distinguish between grants with and without leave to amend)	No	No	Post-*Iqbal* increase in 12(b)(6) grant rates from 21% to 55% (in full) and 24% to 62% (in full or in part) (mean comparisons).
CECIL ET AL., FJC FIRST STUDY (2011)	Order / Party	Excludes counterclaims and prisoner, pro se, and qualified immunity cases	"Near-Census" (data is near-complete population of 12(b)(6) orders in 23 districts)	Yes (controls for district, case type, amended complaint)	Yes (separately codes whether one or more plaintiffs "entirely dismissed" from case)	No	No	No statistically significant post-*Iqbal* increase in 12(b)(6) grant rates, whether in full or in part, with or without leave to amend, or with plaintiff-excluding effect, save in financial instrument cases.

Study Name	Unit of Analysis	Case Type Limitations	Random Sample or Case Census?	Multiple Regression with Covariate Controls?	Isolates Plaintiff-Excluding Dismissals Without Leave to Amend?	Tracks Dis-missals with Leave to Amend?	Accounts for Litigant Selection / Settlement Effects?	Core Findings
CECIL ET AL., FJC SECOND STUDY (2011)	Order / Party	Excludes counterclaims and prisoner, pro se, and qualified immunity cases	"Near-Census" (uses recoded FJC data, as described above)	Yes (controls for district, case type, amended complaint)	Yes (separately codes whether one or plaintiffs "entirely dismissed" from case)	Yes	No	No statistically significant post-*Iqbal* increase in 12(b)(6) grant rates, whether in full or in part or with plaintiff-excluding effect among grants without leave to amend (or with leave but none taken), save in financial instrument cases.
Hatamyar Moore (2012)	Order / Case	Excludes *sua sponte* reviews of prisoner suits, jdx dismissals, fraud cases	Neither (Westlaw search keyed to order's citation of *Conley* or *Twombly* / *Iqbal*)	Yes (controls for circuit, case type, judge type, pro se, class action)	Yes (separately codes whether entire case was "entirely dismissed . . . without leave to amend, or whether some part . . . remained pending")	No	No	Post-*Iqbal* increase in 12(b)(6) grant rates from 46% to 56% (in full) and 74% to 82% (in full or in part) (mean comparisons), increase in rate of grants in full with and without leave to amend (multivariate, but no marginal effects reported), and increase from 52% to 73% among grants in full "entirely dismiss[ing]" non-pro se cases without leave to amend (regression, with marginal effects as calculated and presented in Figure 1, above).
Brescia (2012)	Order	Job and housing discrim. only, excluding orders not directed at allegation "specificity" or involving insurance redlining or Hurricane Katrina-related cases	Neither (Lexis only)	No	No (codes dismissals with prejudice as to at least one claim, but does not gauge plaintiff exclusion)	No	No	Post-*Iqbal* increase in rate of 12(b)(6) grants in full or in part with or without leave to amend in represented cases from 58% to 68% (mean comparisons), but no increase in grants in full or in part without leave to amend (43% versus 43%).

Study Name	Unit of Analysis	Case Type Limitations	Random Sample or Case Census?	Multiple Regression with Covariate Controls?	Isolates Plaintiff-Excluding Dismissals Without Leave to Amend?	Tracks Dismissals with Leave to Amend?	Accounts for Litigant Selection / Settlement Effects?	Core Findings
Gelbach (2012)	Order	Excludes pro se, financial instrument, ADA cases	"Near Census" (uses FJC data, as described above)	No (relies on simple mean-comparisons from FJC study)	No (considers only whether grant dismissed one or more claims without leave to amend or with leave to amend but none taken)	Yes	Yes	Post-*Iqbal* increase in "negatively affected share" of plaintiffs suffering grants in full or in part of 15% (job discrim.), 18% (civil rights), and 22% ("Total Other").
Dodson (2012)	Claim	Excludes fraud cases, jdx dismissals	Neither (Westlaw only)	Yes (claim type, judge party, circuit, pro se, prisoner, published opinion)	No (claim-level focus of study cannot consider plaintiff-excluding effect)	No	No	Post-*Iqbal* increase in individual-claim-level grant rate from 73% to 77% (all cases) and from 81% to 87% (pro se cases only) (mean comparisons), but no statistically significant increase in represented cases.
Hubbard (2013)	Order / Case	"Straddle" cases filed pre-*Twombly* but decided post-*Twombly* / pre-*Iqbal*, excluding certain gov't and other actions, pro se, *in forma pauperis*, fraud, MDL cases	Neither / Census (uses two data sets, one Westlaw-drawn, the other a complete case census)	Yes (controls for circuit, civil rights case)	Yes (tests whether 12(b)(6) case terminations have increased post-*Twombly* as a proportion of all cases)	No	Yes (but only as to plaintiff selection)	No post-*Twombly*/pre-*Iqbal* increase in 12(b)(6) grant or case-termination rates, whether in all cases or civil rights cases (multivariate).

For full bibliographic information, see note 7, above. Unless otherwise noted, all "Core Findings" are reported as marginal effects and are statistically significant at the 95% confidence level using a chi-squared or Fisher's exact test (for mean comparisons) or multivariate regression. Note that Hannon's study reports a single statistically significant multivariate result, finding a 40% increase in the grant rate in civil rights cases, but that finding has apparently been undermined. *See* Hannon, *supra* note 7, at 1838 (noting the 40% increase); Hubbard, *supra* note 7, at 44 n.12 (noting the 40% result's inaccuracy based on a regression specification error).