Article

TWOMBLY AND *IQBAL*'S MEASURE: AN ASSESSMENT OF THE FEDERAL JUDICIAL CENTER'S STUDY OF MOTIONS TO DISMISS

Lonny Hoffman[*]

ABSTRACT

This paper provides the first comprehensive assessment of the Federal Judicial Center's long-anticipated study of motions to dismiss for failure to state a claim after the U.S. Supreme Court's decision in *Ashcroft v. Iqbal*. Three primary assessments are made of the FJC's study. First, the FJC's findings do not indicate that the Court's decisions have had no effect on dismissal practice. To the contrary, the FJC found that after *Iqbal*, a plaintiff was twice as likely to face a motion to dismiss. This sizeable increase in the rate of Rule 12(b)(6) motion activity represents a marked departure from the steady filing rate observed over the last several decades and means, among other consequences, added costs for plaintiffs. Similarly, the data regarding orders resolving dismissal motions demonstrates the consequential impacts of the Court's cases, as in every case type studied there was a higher likelihood after *Iqbal* that a motion to dismiss would be granted. Second, due to the inherent limitations of doing empirical work of this nature, the cases may be having effects that the FJC researchers were unable to detect. Comparing how many motions were filed and granted before *Twombly* with after *Iqbal* does not indicate whether

* George Butler Research Professor of Law, University of Houston Law Center. Stephen Burbank, Aaron Bruhl, Seth Chandler, Kevin Clermont, Elizabeth Demicco, Scott Dodson, David Hoffman, David Kaye, Richard Lempert, Thom Main, Arthur Miller, Alexander Reinert, Joseph Sanders, Charles Silver, Adam Steinman and Tobias Wolf provided comments on an earlier draft of this paper. Special thanks are owed to Joe Cecil for our many communications, for his professionalism, and for his seemingly endless patience with me. The author discussed the paper as an invited Panel Discussant at the Conference on Empirical Legal Studies held at Northwestern University School of Law on November 4-5, 2011. The paper was also presented at the 2011 Southeastern Association of American Law Schools annual meeting in Hilton Head, South Carolina. This project was funded by the University of Houston Law Foundation.

the Court's cases are deterring some claims from being brought, whether they have increased dismissals of complaints on factual sufficiency grounds, or how many meritorious cases have been dismissed as a result of the Court's stricter pleading filter. Finally, the data the FJC researchers gathered may be incomplete, particularly as to the filing rate. As a result, the study may be providing an incomplete picture of actual Rule 12(b)(6) activity.

TABLE OF CONTENTS

INTRODUCTION

The most contentious battleground today in civil litigation concerns the pleading-sufficiency standard in federal court. The Court's bold revision of the pleading test in *Bell Atlantic Corp. v. Twombly*[1] and *Ashcroft v. Iqbal*[2] may have initially raised more questions than answers, but knowledgeable observers recognized immediately that much more was at stake than mere technical requirements for initiating suit. The "cornerstone" of the federal rules, as the architect of the entire structure called it,[3] pleading is the entry point into the system. Deciding how wide or narrow to make the passageway necessarily means deciding how to strike the balance between access to justice, on the one hand, and operational efficiency, on the other.[4]

In *Twombly* and *Iqbal* the Court endorsed more robust filtering at the pleading stage to block certain cases from going further in the litigation process that previously passed unchecked. The defense bar and business community have applauded the use of this less permeable sieve: from their vantage point, intercepting weak claims early in the case—that is, before onerous discovery burdens have to be borne—is vital to the efficient management of civil litigation. Others, including a majority of academics writing on the subject, have criticized the decisions for usurping the Rules Enabling Act process;[5] for adding confusion and unpredictability into the test for pleading sufficiency;[6] for lodging too much discretion in judges,[7] which

---

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

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

[3] Charles E. Clark, *The Influence of Federal Procedural Reform*, 13 LAW & CONTEMP. PROBS., 144, 154 (1948), *available at* http://digitalcommons.law.yale.edu/fss_papers/3253/ ("The cornerstone of the new reform is a system of simple, direct, and unprolonged allegations of claims and defenses by the litigants . . . .").

[4] Prepared Statement of Stephen B. Burbank, Hearing on Whether the Supreme Court has Limited Americans' Access to Court Before the Committee on the Judiciary, United States Senate 2 (Dec. 2, 2009), *available at* www.judiciary.senate.gov/pdf/12-02-09%20Burbank%20Testimony.pdf [hereinafter JUNE 2011 REPORT TO STANDING COMM.] ("The degrees of particularization and persuasiveness of a complaint's allegations that a system requires implicate the ability of putative plaintiffs to pursue adjudication of disputes on the merits . . . . They thus also implicate the ability of those who have been injured to use litigation in order to secure compensation, and the ability of government to use private litigation for that purpose (i.e., in place of social insurance), and for the enforcement of social norms (i.e., in place of administrative enforcement)."); *see also* Lonny S. Hoffman, *Burn Up the Chaff with Unquenchable Fire: What Two Doctrinal Intersections Can Teach Us About Judicial Power Over Pleadings*, 88 B.U. L. REV. 1217, 1222 (2008) (suggesting that *Twombly* may "mark a fundamental change in where courts strike the balance between access and efficiency"); *see also* Phillips v. Cnty. of Allegheny, 515 F.3d 224, 230 (3d Cir. 2008) ("Few issues in civil procedure jurisprudence are more significant than pleading standards, which are the key that opens access to courts.").

[5] *See, e.g.,* Burbank, *supra* note 4 at 15-16 ("In initiating change through its power to decide cases and controversies, however, the Court was forced to forego the informational, participatory and other benefits that the rulemaking process affords."); Kevin M. Clermont & Stephen C. Yeazell, *Inventing Tests, Destabilizing Decisions*, 95 IOWA L. REV. 821, 850 (2010) ("The rulemaking bodies should have hosted that discussion. *Twombly* and *Iqbal* short-circuited any such discussion. These cases worked their reform by a process—adjudication—that is hardly the preferred path to design change.").

[6] *See, e.g.*, Clermont & Yeazell, *supra* note 5, at 823 ("By inventing a new and foggy test for the threshold stage of every lawsuit, [*Twombly* and *Iqbal*] have destabilized the entire system of civil litigation."); A. Benjamin Spencer, *Understanding Pleading Doctrine*, 108 MICH. L. REV. 1, 9 (2009) (noting that "*Twombly*'s ultimate message regarding pleading standards is unclear"); Hoffman, *supra* note 4, at 1258 ("[A]mbiguity in the standard for determining which cases will receive greater scrutiny means imposing

fosters inconsistency and arbitrariness;[8] and for turning on its head the basic presumption of modern procedural law for resolving cases on their merits.[9]

Prompted by these criticisms, several bills were introduced in Congress that would have reversed the Court's decisions. From the start, however, these bills have lacked political traction. One important reason (though not the only one) for this is that they were opposed by the Judicial Conference, which has taken the position that legislators should allow judicial rulemakers to study the empirical effects of the decisions and then, through the rule-making process, decide what corrective measures, if any, are needed. Congress does not always follow the advice of the Judicial Conference, of course, but with most Republicans already predisposed against the legislation, the Conference's opposition suggests that the prospects for legislative reform are dim.

In the absence of any meaningful possibility that Congress will act, the task has fallen to judicial rulemakers to decide whether and how to respond to the U.S. Supreme Court's decisions. However, while rulemakers have heard all of the theoretical arguments against *Twombly* and *Iqbal*, they have not been persuaded that amendments to the pleading rules are necessary to counteract the Court's decisions, especially without convincing empirical evidence that

---

additional costs on everyone, thus carrying serious practical and social consequences.").

[7] *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, 29 (2010) ("Although judicial discretion normally is to be applauded, it should be constrained in the context of a threshold motion theoretically addressed solely to the notice-giving quality and legal sufficiency of the complaint."); Howard M. Wasserman, Iqbal, *Procedural Mismatches, and Civil Rights Litigation*, 14 LEWIS & CLARK L. REV. 157, 177 (2010) (noting that under *Twombly* and *Iqbal*, courts "enjoy broad discretion to parse the complaint and individual allegations and to screen aggressively for a story that resonates with them"); Elizabeth Thornburg, *Law, Facts, and Power*, 114 PENN ST. L. REV. PENN STATIM 1, 10 (2010), *available at* www.pennstatelawreview.org/114/114%20Penn%20Statim%201.pdf (describing the doctrinal test in the Court's cases as a "magic trick" that has "privileged judges over juries, appellate judges over trial judges, and put the Court firmly at the top of the heap").

[8] *See, e.g.*, Miller, *supra* note 7, at 30 (observing that "inconsistent rulings on virtually identical complaints may well be based on individual judges having quite different subjective views of what allegations are plausible"); Hoffman, *supra* note 4, at 1258 ("Plausibility is not only an uncertain standard by which to measure when greater scrutiny is warranted, but it also, and more mischievously, invites a free-wheeling judicial judgment as to the legitimacy of claims. That should give cause for concern, especially given the anti-plaintiff influence [of *Twombly*] . . . and courts that want to exercise their newly-minted authority to dispose of those cases they perceive to be unwelcome will not miss it.").

[9] *See, e.g.*, Miller, *supra* note 7, at 29-30 ("[The *Twombly/Iqbal*] process is uncomfortably close to a weighing of the evidence and an invasion of the jury's domain, suggesting that the Court's decisions represent a potentially significant change in the division of functions between judge and jury. In other words, a trial-like scrutiny of the merits is being shifted to an extremely early point in the pretrial phase."); Kevin M. Clermont, *Three Myths About* Twombly-Iqbal, 45 WAKE FOREST L. REV. 1337, 1348 (2010) ("*Twombly-Iqbal* calls for a judge to weigh factual convincingness without any evidential basis and with few procedural protections. Such a practice, in the absence of emergency or other special circumstances, offends our fundamental procedural principles."); Suja A. Thomas, *The New Summary Judgment Motion: The Motion to Dismiss Under* Iqbal *and* Twombly, 14 LEWIS & CLARK L. REV. 15, 41 (2010) (arguing that the "motion to dismiss is now the new summary judgment motion, in standard and possibly effect" and that "[t]hese similarities of the standards and possibly the effects of the motions call into question whether Iqbal and Twombly were decided properly").

the cases are impacting dismissal practice.[10] The key modifier in that last sentence was *convincing* empirical evidence. Some prior studies suggested *Twombly* and *Iqbal* were making it harder for at least some plaintiffs to overcome the new pleading barrier, but it was not clear that these studies—which were drawn from the selected opinions found in electronic databases—were representative of dismissal practices generally. Accordingly, the rules committees commissioned the Federal Judicial Center to undertake a more comprehensive examination of dismissal activity.

Released in March 2011,[11] the FJC's study featured a headline that did not square with academic predictions and the prior empirical research: *Twombly* and *Iqbal* were not having much effect on dismissal practices or outcomes, after all.[12] The critical point that the FJC researchers emphasized to readers was that they found no "statistically significant" increase in the likelihood that a motion to dismiss would be granted after *Iqbal* (except for one outlier case category). The study's key finding was summarized this way:

> [W]e found a statistically significant increase in the rate at which motions to dismiss for failure to state a claim were granted only in cases challenging financial instruments. . . . We found no increase in the rate at which motions to dismiss were granted, with or without opportunity to amend, in other types of cases.[13]

---

[10] *See, e.g.,* MARK R. KRAVITZ, ET AL., ADVISORY COMM. ON FED. R. OF CIV. P., REPORT OF THE CIVIL RULES ADVISORY COMMITTEE TO THE STANDING COMMITTEE ON RULES OF PRACTICE AND PROCEDURE 9 (Dec. 6, 2010) *in* 1 AGENDA MATERIALS FROM THE COMM. ON RULES OF PRACTICE & PROCEDURE, JAN. 2011, at 102, *available at* http://www.uscourts.gov/RulesAndPolicies/FederalRulemaking/ResearchingRules/AgendaBooks.aspx [hereinafter JAN. 2011 REPORT TO STANDING COMM.] ("[Since *Twombly* and *Iqbal*,] pleading standards have been moved from a continuing but inactive status on the agenda to active consideration. Active consideration does not imply a plan for imminent rules proposals. To the contrary, it is better to wait patiently while lower courts work through the ways in which pleading practice should be adjusted to meet the concerns expressed by the Supreme Court."); MARK R. KRAVITZ, ET AL., ADVISORY COMM. ON FED. R. OF CIV. P., REPORT OF THE CIVIL RULES ADVISORY COMMITTEE TO THE STANDING COMMITTEE ON RULES OF PRACTICE AND PROCEDURE 54 (May 2, 2011) *in* AGENDA MATERIALS FROM THE COMM. ON RULES OF PRACTICE & PROCEDURE, JUNE 2011, at 217, *available at* http://www.uscourts.gov/RulesAndPolicies/FederalRulemaking/ResearchingRules/AgendaBooks.aspx [hereinafter JUNE 2011 REPORT TO STANDING COMM.] ("The [Civil Rules Committee's] approach to pleading practice remains what it has been since 2007. The Committee will closely monitor developing practice, it will encourage and heed further rigorous empirical work, and it will listen carefully to the voices of bench, bar, and academy. Procedural ferment is exciting, but it does not justify an excited response."). It should be said that the Committees are still actively considering other reforms short of pleading rule reform, such as discovery rule amendments. Discovery rule reforms may be responsive to some concerns raised by *Twombly* and *Iqbal*, at least for some claimants, but it is not clear that they would be adequate to overcome all concerns. *See supra* text accompanying notes 5-9.

[11] JOE S. CECIL, ET AL., FED. JUDICIAL CTR., MOTION 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.fjc.gov/public/pdf.nsf/lookup/motioniqbal2.pdf/$file/motioniqbal2.pdf.

[12] *See id.* at vii.

[13] *Id.* at 21.

With the evidence apparently showing that concerns about *Twombly* and *Iqbal* were premature, the FJC's work is now being cited as powerful support for the case against pleading rule reform.[14]

The problem with this interpretation of the study's findings is that it is greatly, if unintentionally, misleading. By summarily announcing that the observed increases were not statistically significant, but not explaining what that technical terminology means (and, as importantly, what it does not mean), the study confuses readers into thinking that it demonstrated the Court's decisions had no impact on dismissal practice. The study proved no such thing. To be precise, the researchers failed to find strong statistical evidence that the decisions were causally related to the changes that were found to have occurred. The failure to prove the decisions had an effect is not the same as proof that the decisions had no effect. At most, when detected effects are not statistically significant, it tells us only that we cannot rule out the possibility that they may have been the result of random chance and not a real association.[15]

Moreover, saying the observed effects were not statistically significant certainly does not mean that the researchers observed no effects. However, because of the focus on statistical significance, and the occasional use of imprecise language, the study may lead at least some readers to overlook the considerable changes in dismissal practices and outcomes the researchers did observe in comparing dismissal motions and orders before *Twombly* with motions and orders after *Iqbal*. For starters, the rate of dismissal motions that were filed increased substantially. After *Iqbal*, a plaintiff was twice as likely to face a motion to dismiss as compared with the period before *Twombly*, a marked increase in the rate of Rule 12(b)(6) motion activity from the steady filing rate observed over the last several decades. As for dismissal orders, the FJC found that in every case category that was examined there were more orders granting dismissal after *Iqbal* than there were before *Twombly*, both with and without prejudice. Most importantly, in every case category examined it was more likely that a motion to dismiss would be granted. The higher success rate extended only to motions granted with leave to amend, but the researchers also found that the movant's success rate was significantly higher when the plaintiff had previously had a chance to amend her complaint. Not only is this to be expected, but it also suggests reasons to be concerned that a dismissal granted pursuant to *Twombly* and *Iqbal* on the basis that allegations are "conclusory," "implausible," or both, is not easily cured merely by allowing an opportunity to amend allegations that have already been made.

---

[14] At the June 2011 meeting of the Standing Committee (the most recent meeting of rulemakers, as of this writing), the FJC study figured prominently in the report submitted to the Standing Committee by the Civil Rules Advisory Committee. *See* JUNE 2011 REPORT TO STANDING COMM., *supra* note 10, at 215-16 ("[T]he lower-court decisions may suggest that not much has changed in actual practice. That hypothesis finds support in the first detailed study done by the Federal Judicial Center. . . ."). *See also id.* at 216 ("The FJC study—and the promise of its next study—combines with the review of judicial decisions to suggest there is no urgent need for immediate action on pleading standards. The courts are still sorting things out. There is reason to hope that the common-law process of responding to and refining the Supreme Court's invitation to reconsider pleading practices will arrive at good practices.").

[15] *See infra* Part III.

Given these sizeable differences in dismissal practices and outcomes from the period before *Twombly* to the period after *Iqbal*, what readers needed to understand was what it means to say that the observed differences were not statistically significant. Broadly stated, my argument is that the study failed to help readers answer that critical question. Rather than summarily announcing that the detected effects were not statistically significant, the researchers should have aided transparency and understanding by explicitly discussing how to interpret the study's results. The need for clarity was particularly acute with this study because most of its readers likely will not possess background training in statistics. Indeed, the rules committee that commissioned the study has shown that they rely heavily on the researchers' description of the study's findings in making their policy decisions.[16]

One important way that researchers could have aided transparency was by pointing out that demanding a high degree of confidence, as the researchers did, before concluding that chance can be ruled out as an explanation carries real consequences for policymaking. The primary consequence is that it makes an implicit (but not obvious) policy judgment that our highest concern should be to avoid being gullible—that is, thinking that the jump in post-*Iqbal* dismissal orders was caused by the Court's decisions when, in fact, it was not. However, the cost of worrying so much about this type of error is that it makes it more likely that policymakers will make a different, equally worrisome mistake—that is, thinking the decisions are having no effect when, in fact, they are. Both kinds of errors matter, and, at a minimum, the researchers should have clarified that they were measuring only the likelihood of the first kind of error. Better still, the researchers should have tried to estimate the appropriate level of statistical significance by taking into consideration the costs of both types of errors.

In addition to clarifying the cost of setting an inappropriately high threshold for statistical significance, the FJC researchers should also have made clear to its general readers that statistical testing is an art, not a science, or as Roger Kirk has put it, not some "objective, scientific procedure" for which there is no "element of subjectivity."[17] To this end, they should have pointed out that the observed effects might not have been statistically significant for a number of reasons, including that the models used were incorrectly constructed, or that the wrong statistical test was employed or, as noted above, that an inappropriately high level of statistical significance was used. Any of these possibilities, which I discuss in greater detail in Part III, could reasonably explain why the differences they observed between the pre-

---

[16] *See, e.g.*, Draft Minutes of Civil Rules Advisory Committee Meeting, April 4-5, 2011, *in* AGENDA MATERIALS FROM THE COMM. ON RULES OF PRACTICE & PROCEDURE, JUNE 2011, at 300-06, *available at* http://www.uscourts.gov/RulesAndPolicies/FederalRulemaking/ResearchingRules/AgendaBooks.aspx (reflecting the extended discussion of the FJC's *Iqbal* Study at the April 2011 meeting of the Advisory Committee on Civil Rules). *See also* THOMAS E. WILLGING, FED. JUDICIAL CTR., USE OF RULE 12(b)(6) IN TWO FEDERAL DISTRICT COURTS 3 (1989), *available at* http://www.fjc.gov/library/fjc_catalog.nsf (noting that, based on study findings, the Advisory Committee on Civil Rules decided not to act on a prior proposal by Professor Carrington, then the reporter for the Civil Rules Committee, to abrogate Rule 12(b)(6)).

[17] Roger E. Kirk, *Promoting Good Statistical Practices: Some Suggestions*, 61 EDUC. & PSYCHOL. MEASUREMENT 213, 214 (2001).

Twombly and post-*Iqbal* period did not reach their defined level of statistical significance. With this awareness of the limits of statistical testing, readers might have better understood how to interpret the study's limited results.

The paper is organized as follows. Part I begins with a brief summary and discussion of the study's findings. Part II examines and critiques the study's findings regarding differences in the rate at which defendants filed dismissal motions in the pre-*Twombly* and post-*Iqbal* periods. Part III, which focuses on the FJC's findings regarding dispositions of dismissal motions, provides the primary critiques of the study. My argument is that the researchers could have aided transparency and understanding by (1) clarifying what a "no significance" finding means; (2) discussing the cost of setting an inappropriately high threshold for statistical significance; and (3) raising with readers plausible reasons why the differences they observed in dismissal outcomes might not have reached their defined level of statistical significance.[18] By failing to clarify the results in these ways, the study unintentionally confused readers into thinking that it proved that *Twombly* and *Iqbal* were not responsible for the dismissal increases observed and made it more likely that the considerable evidence of effects that the study was able to observe would be overlooked.

Part IV then argues that there may be other effects the cases are having that the FJC researchers would not have been able to observe given the study's design. Finally, Part V explores the possibility that the data the FJC researchers gathered may be incomplete, particularly as to the filing rate. As a result, the study may be providing an incomplete picture of actual Rule 12(b)(6) activity. I end with a short conclusion and appeal that readers of this vitally important and influential study, and policymakers in particular, should carefully reassess the study's findings in light of the criticisms and assessments offered.

I. WHAT THE FJC STUDIED AND FOUND

A. Study Design

Several attempts have been made to study the effects of the Court's decisions.[19] Due to resource and informational constraints, most studies have focused exclusively on opinions found in electronic databases, such as Westlaw. The key difference between the FJC's study and prior empirical studies of *Twombly* and *Iqbal* is that the FJC looked at all dismissal activity,

---

[18] As of this writing, the dataset the FJC examined was not publicly available. Access to federal court records for academic research purposes is often limited by the fees associated with use of the Public Access to Court Electronic Records (PACER) system in federal court, as well as functional constraints. *See generally* Lynn M. LoPucki, *Court-System Transparency*, 94 IOWA L. REV. 481 (2009).

[19] *See* Patricia W. Hatamyar, *The Tao of Pleading: Do* Twombly *and* Iqbal *Matter Empirically?*, 59 AM. U. L. REV. 553, 584-633 (2010); Joseph A. Seiner, *Pleading Disability*, 51 B.C. L. REV. 95, 116-29 (2010); Joseph A. Seiner, *The Trouble with* Twombly*: A Proposed Pleading Standard for Employment Discrimination Cases*, 2009 U. ILL. L. REV. 1011, 1027-32; 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, 1828-46 (2008).

whether or not the orders appeared in Westlaw. Looking at actual activity in the district courts is a more comprehensive approach to gathering the sought-after data than limiting one's data collection efforts to only those opinions published on Westlaw. Opinions published on Westlaw constitute less than all district court decisions, and the concern is that relying on published opinions may not be representative of all dismissal orders.[20]

More precisely, the FJC study compared motion activity in 23 federal district courts before *Twombly* and after *Iqbal*.[21] The database was generated by relying on codes entered by the court clerks of the individual districts into a file management system used by the federal courts called Case Management/Electronic Case Filings, or CM/ECF as it is known. The CM/ECF codes entered by the clerks relate to motions filed by lawyers and orders issued by judges in individual cases.[22] The basic construct of the FJC study was to compare the pre-*Twombly* and post-*Iqbal* rate at which motions to dismiss for failure to state a claim were filed and the success rate of these motions.

B. Findings Regarding the Filing Rate

Doing a straightforward comparison of filing rates, the FJC found that, overall, motions to dismiss for failure to state a claim were brought more often after *Iqbal* (that is, in cases filed from October 2009 to June 2010) than before *Twombly* (cases filed from October 2005 through June 2006). In the earlier period, the FJC found that defendants filed motions to dismiss for failure to state a claim in 4% of all civil cases. After *Iqbal*, the rate increased to 6.2%. The overall increase was reported as statistically significant at the .01 level.[23] It was also reported that there was a statistically significant increase in the filing rate for all individual categories of cases except civil rights cases.[24] Table 1 from the FJC study illustrates the filing-rate findings as to the six main

---

[20] CECIL, ET AL., *supra* note 11, at 2, 37 n.47. However, recent work by Patricia Hatamyar Moore shows that orders granting 12(b)(6) motions are just as likely to be "published" in Westlaw as orders denying 12(b)(6) motions. *See* Patricia W. Hatamyar, *An Updated Quantitative Study of* Iqbal*'s Impact on 12(b)(6) Motions*, 45 U. RICH. L. REV. (Aug. 8, 2011) (forthcoming), at 1-4 *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1883650 (updating her prior study that had examined *Iqbal*'s effects by limiting her database as the FJC did, and finding a high degree of consistency between her findings and those in the FJC's *Iqbal* study).

[21] CECIL, ET AL., *supra* note 11, at 5. The twenty-three federal district courts in the study primarily included the two districts in each of the eleven circuits with the largest number of case filings in 2009, plus the U.S. District Court for the District of Columbia. *Id.* Because it was not possible to collect the data in some of the largest districts, they had to be excluded. *Id.* Together, the twenty-three district courts account for just over half of all civil cases filed in 2009. *Id.*

[22] *Id.*

[23] *Id.* at 8-9 & tbl.1.

[24] In the Civil Rights category, the increase went from 9.7% to 10.1%, but, as the researchers noted, it did not reach statistical significance at the .05 level. *Id.* at 8-9 and Table 1. The statistical significance story gets a bit more nuanced here, however. In the Civil Rights category, three-fourths of the cases were non-prisoner civil rights alleging constitutional violations under 42 U.S.C. §1983. This subset of all Civil Rights cases, the researchers noted, "showed a statistically significant increase ($p \leq 0.05$) in the likelihood that a motion to dismiss for failure to state a claim would be filed, up from 10.5% of cases in 2006 to 12.4% of cases in 2010." *Id.*

case categories:

Table 1: Percentage of Civil Cases with a Motion to Dismiss for Failure to State a Claim Filed Within 90 Days of the Filing of the Case (Excluding Prisoner and Pro Se Cases)

	2005–2006 Percentage (and Number) of Cases	**2009–2010 Percentage (and Number) of Cases**	**Difference**
Total	4.0% (49,443)	6.2% (52,925)	+2.2%*
Contract	5.6% (8,651)	8.3% (9,139)	+2.7%*
Torts	2.3% (10,604)	4.1% (9,947)	+1.8%*
Employment Discrimination	6.9% (3,795)	9.0% (3,871)	+2.1%*
Civil Rights	9.7% (4,214)	10.1% (4,976)	+0.4%
Financial Instrument	4.3% (1,524)	9.6% (4,790)	+5.3%*
Other	2.5% (20,657)	4.1% (20,202)	+1.6%*

*$p < 0.01$.

Table 1 is reprinted from FJC Study at 9.

After regression analysis, the results of these straightforward comparisons were confirmed. The likelihood that a motion to dismiss would be filed in any individual case increased after *Iqbal*, as compared with a baseline that was constructed to measure changes in the filing rate over time and across different kinds of cases. In the post-*Iqbal* period it was twice as likely that a plaintiff would face a motion to dismiss.[25] The filing rate also trended up, on a monthly basis, in the post-*Iqbal* period, in contrast to the monthly trend line in the 2005-06 time period, which remained essentially flat.[26]

C. Findings Regarding Dispositions of Motions

In addition to looking at filings, the FJC also examined how often movants were successful in obtaining dismissal. The study found that there was an increase in the number of orders granting dismissal in the post-*Iqbal* period, both with and without prejudice to amend, both overall and in every case category examined. Although the researchers did not focus attention on the

[25] *Id.* at 9-10 & tbl.2.

[26] *Id.* at 10-11 (noting that "the percentage of cases with one or more motions to dismiss for failure to state a claim was higher in each month of 2009–2010 than in each month of 2005–2006"; and "in 2009–2010 there appeared to be a modest increase over time in the percentage of cases with such motions"). *See also id.* at 11 fig.1.

increases, they are reflected in Table 4 of the study. Below is an excerpt from the top part of Table 4, which shows the overall increase in dismissal orders:

Table 4: Outcome of Motions to Dismiss for Failure to State a Claim

	Action on Motion	2006	No. of Orders	2010	No. of Orders	Difference
Total	Denied	34.1%	(239)	25.0%	(305)	
	Granted All or Some Relief	65.9%	(461)	75.0%	(916)	+9.1%*
	With Amendment	20.9%	(146)	35.3%	(431)	+14.4%†
	Without Amendment	45.0%	(315)	39.7%	(485)	-5.3%

* $p \leq 0.01$, relative to the likelihood that the motion will be denied.
† $p \leq 0.05$, relative to the likelihood that the motion will be granted without leave to amend.

Table 4 is reprinted from FJC Study at 14.

The total number of orders granting dismissal for each case category examined, both with and without leave to amend, increased in the post-*Iqbal* period.

There was an increase in case filings in the later period, but it was not nearly as large as the increase in dismissal orders granted. Civil case filings increased only 7% in the 23 federal district courts from which the FJC's data was drawn. This change can be contrasted with the percentage increase in the total number of orders granting dismissal, as depicted in Figure 1.

Figure 1: Percentage Rise after *Iqbal* in Total Number of Orders Granting Dismissal with Leave to Amend

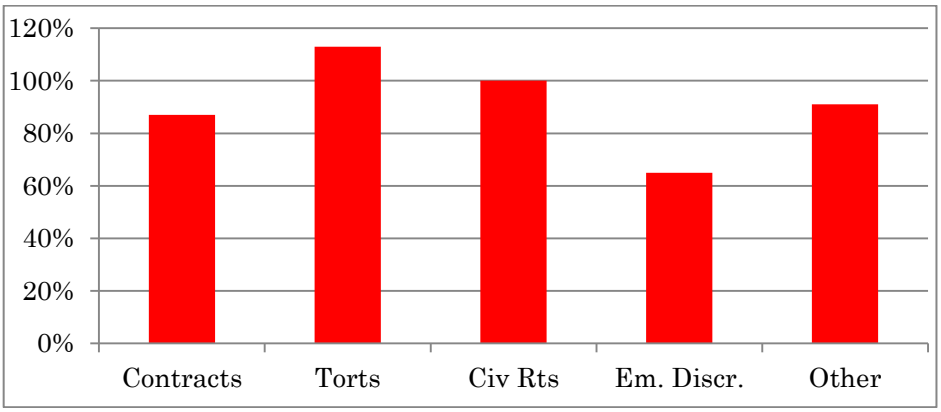

Figure 1: Columns depict the percentage rise in the total number of orders granting a defendant's motion to dismiss for failure to state a claim with leave to amend the complaint. Results are calculated from data reported in FJC Study at 12-14 and Table 4. Excluded are the findings for Financial Instruments cases.

The FJC study also found that, on average, defendants were more successful in bringing motions to dismiss since *Iqbal*. Across all cases, the overall grant rate went from 66% in the earlier period to 75% in the latter period, as the excerpted portion of Table 4, above, shows. More precisely, the data reflect that in the three largest case categories (Other, Financial Instruments and Civil Rights), it was much more likely after *Iqbal* that a court would grant a motion to dismiss with leave to amend. The rate at which motions to dismiss were granted with leave to amend increased 12.8, 30.5 and 11.7 percentage points, respectively. The remaining three categories (Contract, Torts and Employment Discrimination) show smaller but still clearly increasing grant rates. I have more graphically illustrated in Figure 2 below the magnitude of increase in the percentage of orders granting dismissal with leave to amend after *Iqbal* for every case category examined (excluding the Financial Instruments cases).

Figure 2: Percentage of Orders Granting Dismissal with Leave to Amend

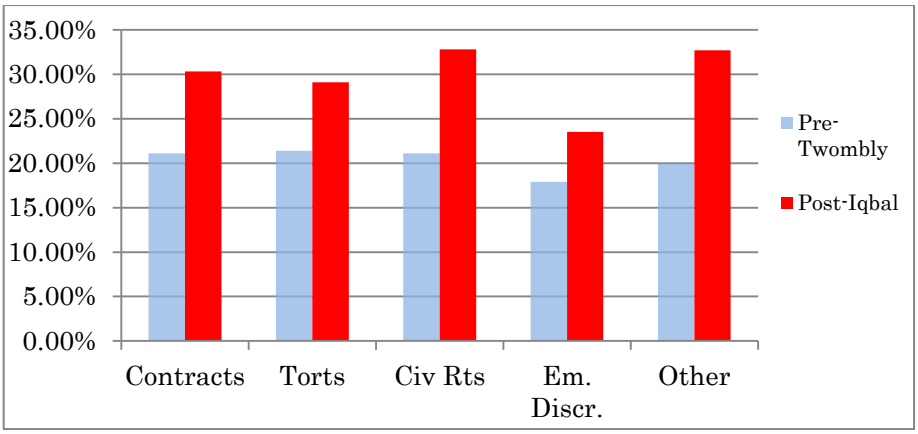

Figure 2: Clustered columns depict the percentage of orders granting a defendant's motion to dismiss for failure to state a claim with an opportunity to amend the complaint. All data are drawn from the FJC Study (at 12-14 and Table 4). Excluded are the findings for Financial Instruments cases.

Even more starkly put, there were substantial rises in the grant rate, including a 64% rise in cases in the Other category and a 55% rise for Civil Rights cases. Figure 3 graphically depicts the percentage rises for all case categories:

Figure 3: Percentage Rise after *Iqbal* in Orders Granting Dismissal With Leave to Amend

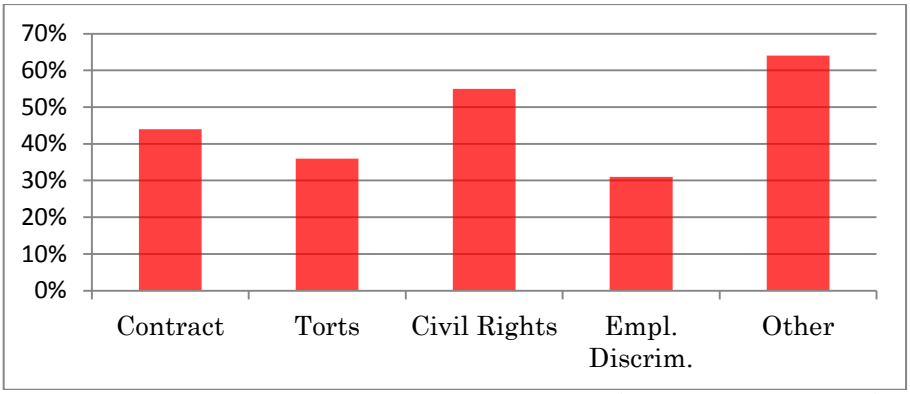

Figure 3: Columns depict the percentage rise in orders granting a defendant's motion to dismiss for failure to state a claim with an opportunity to amend the complaint. Results are calculated from data reported in FJC Study at 12-14 and Table 4. Excluded are the findings for Financial Instruments cases.

Notwithstanding these sizeable percentage increases in the likelihood that a motion to dismiss would be granted after *Iqbal*, the FJC researchers urged caution in interpreting the grant rate data for two reasons.

The first reason for not reading too much into this finding, the researchers observed, was that the higher grant rate was only for grants with leave to amend. This is how they put it:

> An important reason for caution in interpreting these differences is that in 2010, orders granting motions to dismiss were far more likely to allow the plaintiff to amend the complaint, leaving open the possibility that the plaintiff might cure the defect in the complaint and the case might proceed to discovery.[27]

In other words, the researchers suggested that dismissals with leave to amend may be less worrisome than outright dismissals with prejudice. Moreover, they continued, an additional and even more critical reason to be cautious in interpreting even the data regarding dismissals with leave to amend is that, except for financial-instrument cases, the increases were not "statistically significant" for any other case category.[28] The study's discussion section put it this way:

> After controlling for identifiable effects unrelated to the Supreme Court decisions, such as differences in caseload across individual districts, we found a statistically significant increase in the rate at which motions to dismiss for failure to state a claim were granted only in cases challenging financial

[27] CECIL, ET AL., *supra* note 11, at 13.
[28] *See id.* at 13-14 & tbl.4, 21.

> instruments. . . . We found no increase in the rate at which motions to dismiss were granted, with or without opportunity to amend, in other types of cases.[29]

The researchers did not explain, however, what it meant for the increases they observed in the dismissal rate not to be statistically significant.

Several other grant-rate findings in the study also bear brief mention. The researchers found that a motion that sought dismissal of an amended complaint had a better chance of being granted than if dismissal was sought of an original complaint.[30] Also observed were differences from one district to another as to orders granting dismissal motions, both with and without leave to amend.[31] Finally, the researchers reported no difference after *Iqbal* in how speedily cases were terminated after an order of dismissal. They noted that "if the district courts were interpreting *Twombly* and *Iqbal* to significantly foreclose the opportunity for further litigation in the case, we would expect to see an increase in cases terminated soon after the order." However, "we found no statistically significant increase in 2010 in the percentage of cases terminated in 30 days, 60 days, or 90 days after the order granting the motion.[32]

II. INTERPRETING STUDY TO FIND LITTLE EVIDENCE OF *TWOMBLY* AND *IQBAL*'S EFFECTS MISREADS KEY FINDINGS OF CASES' IMPACT: THE EVIDENCE REGARDING FILINGS

This Part examines more closely the study's findings regarding the frequency with which motions to dismiss were brought. The next part (Part III) turns to dispositions of dismissal motions.

The FJC's study confirms early predictions that *Twombly* and *Iqbal* would incentivize defendants to challenge the sufficiency of the plaintiff's complaint more frequently. The researchers found a 50% increase from before *Twombly* to after *Iqbal* in the rate at which motions to dismiss for failure to state a claim were filed. Regression analysis to control for differences across federal districts and across types of cases confirmed the straightforward findings: after *Iqbal*, a plaintiff was twice as likely to face a motion to dismiss.[33] This sizeable increase in the rate of Rule 12(b)(6) motion activity represents a marked departure from the steady filing rate observed over the last several decades.[34] Recall further that the FJC also found an increasing month-to-month trend line in the post-*Iqbal* period, providing some (though perhaps weak) evidence to suggest that the filing rate may continue to rise over time.[35]

Of course, the preceding discussion assumes the filing-activity levels the FJC found are accurate. In Part V, below, I discuss the possibility that the data the FJC researchers gathered may be incomplete, particularly as to the filing rate. If instead of the 4% pre-*Twombly* rate the FJC reported, the actual filing

---

[29] *Id.* at 21 (referencing results reported in Table 4).
[30] *See id.* at 19.
[31] *Id.* at 18-19 & tbl.8.
[32] *Id.* at 16 & tbl.6.
[33] *Id.* at 10 & tbl.2.
[34] *See infra* notes 75-78.
[35] CECIL, ET AL., *supra* note 11, at 10-11 & fig.1.

rate before *Twombly* was closer to 13-15% as prior FJC studies suggest it may actually have been, then applying the same 50% increase in the filing rate would mean that defendants after *Iqbal* may be filing motions, on average, in roughly one out of every five cases. Moreover, keep in mind that the above figures refer to the average filing rate across all cases. The FJC's 2011 study observed filing rates (both before *Twombly* and after *Iqbal*) for employment discrimination and other civil rights (non-prisoner) cases that were well above the average rate.[36] Prior empirical study of motions to dismiss similarly recorded higher filing rates for these two important case types but at even higher rates than the FJC's 2011 study found.[37]

That more motions are being filed carries real consequences for litigants. It means added costs for those who have to gather additional information either in anticipation of or in response to these motions. It also means added costs in having to defend against these more frequently filed motions, even those that ultimately are unsuccessful.[38] Writing before the FJC's study, Arthur Miller anticipated that "federal courts will be required to devote much more time to evaluating factual allegations than in the past—time that might be better spent appraising the merits of a well-developed record presented at summary judgment or trial, especially with regard to uncomplicated matters."[39] Moreover, none of these cost calculations take into account that some litigants will be unable to bear the additional expenses, or will lack access to the information sought, and so either will be deterred from bringing suit or unable to stave off dismissal. All of these are additional consequences that flow directly from the greater willingness of defendants to bring motions to dismiss (but they are consequences that were invisible to the FJC researchers who were not looking for those effects, as discussed further in Part IV).[40]

---

[36] *Id.* at 9 tbl.1 (reporting, *inter alia*, a pre-*Twombly* filing rate of 4% for all cases, as compared with 6.9% and 9.7% for employment discrimination and other civil rights (non-prisoner) cases, respectively).

[37] WILLGING, *supra* note 16, at 12 tbl.4 (reporting 9% filing rate for employment discrimination cases and 12% filing rate for civil rights (non-prisoner) cases). *See infra* Part V.A for a discussion of the 1989 study.

[38] *See* Clermont & Yeazell, *supra* note 5, at 840-41 (predicting an increase, after *Iqbal*, in the number of motions to dismiss that are filed and observing that "many plaintiffs will bear the expensive burden of these motions, even if the motions fail").

[39] Miller, *supra* note 7, at 41-42.

[40] Moreover, as Jonah Gelbach has argued, comparisons of dismissal rates before and after the Court's decisions in *Twombly* and *Iqbal* will tend to understate their impact because they ignore party-selection effects. Jonah B. Gelbach, Note, *Locking the Doors to Discovery? Conceptual Challenges in and Empirical Results for Assessing the Effects of* Twombly *and* Iqbal *on Access to Discovery*, 121 YALE L.J. (Dec. 19, 2011) (forthcoming) at 21-36, *available at* http://ssrn.com/abstract=1957363. Gelbach calculates the impact of the decisions in preventing cases from reaching the discovery stage by separately identifying cases in which a motion to dismiss would have been filed both before and after *Twombly* ("non-selection cases") and those in which defendants would only have filed a motion to dismiss in the post-*Iqbal* regime ("defendant selection cases"). *Id.* at 49-65. From this, Gelbach concludes that, excluding civil rights and employment discrimination cases, the Court's decisions in *Twombly* and *Iqbal* caused no fewer than approximately one in five of cases overall to fail to reach discovery and that nearly one-third of cases in which a dismissal motion is now filed will be dismissed because of *Twombly* and *Iqbal*'s heightened pleading standard. *Id.* at 63.

III. INTERPRETING STUDY TO FIND LITTLE EVIDENCE OF *TWOMBLY* AND *IQBAL*'S EFFECTS MISREADS KEY FINDINGS OF CASES' IMPACT: THE EVIDENCE REGARDING DISPOSITIONS

In this Part the focus shifts from the filing rate to the study's findings regarding dispositions of motions. As noted above, the FJC found both a higher absolute number of orders granting dismissal in the post-*Iqbal* period and a higher likelihood that motions would be granted in every case category examined. We have also seen, however, that the researchers focused little attention on the absolute increase in orders and that, as far as the grant-rate findings are concerned, the researchers emphasized that the post-*Iqbal* increases were not statistically significant in any case category except for Financial Instruments cases.[41]

Whatever the benefits of a well-constructed study, empirical research can also confound thinking if the chosen methodology is unsound or if even adequately collected findings are not communicated clearly.[42] By emphasizing that none of the dismissal increases in the other case categories were statistically significant, the FJC's study leads readers to assume that the study proved *Twombly* and *Iqbal* were not responsible for the higher number and rate of dismissals, as well as to overlook the effects that were observed. The researchers should have communicated the study's results more clearly. More precisely, I argue that the researchers could have aided transparency and understanding by (1) clarifying what a finding of no significance means; (2) discussing the cost of setting an inappropriately high threshold for statistical significance; and (3) raising plausible reasons why the observed differences might not have reached the level of statistical significance they selected.

A. The Researchers Should Have Clarified What a Finding of No Significance Means

The FJC researchers could have greatly aided understanding of the study's results by clarifying what a finding of no significance means. The failure to keep clear the limits of null hypothesis statistical testing has been the source of countless problems in the social sciences and biomedical fields.[43] To better understand what a finding of statistical significance means—and so to explain what the all-powerful p-value denotes (and what it does not denote)—it is helpful to clarify why calculations of statistical significance are made.

Step outside the field of law and consider a biomedical researcher who is interested in determining whether a certain drug has an effect on people. The work the FJC did in collecting data on dismissal rates may not have been strictly analogous but the exercise helps to provide a better appreciation for statistical testing. Assume that a researcher conducts an experiment in which

---

[41] CECIL, ET AL., *supra* note 11, at 21.

[42] *See, e.g.*, Lee Epstein & Gary King, *Exchange: Empirical Research and the Goals of Legal Scholarship, The Rules of Inference*, 69 U. CHI. L. REV. 1, 1-18 (2002); Lee Epstein, Andrew D. Martin & Matthew M. Schneider, *On the Effective Communication of the Results of Empirical Studies, Part I*, 59 VAND. L. REV. 1811, 1812-15 (2006).

[43] *See* STEPHEN T. ZILIAK AND DEIRDRE N. MCCLOSKEY, THE CULT OF STATISTICAL SIGNIFICANCE: HOW THE STANDARD ERROR COSTS US JOBS, JUSTICE, AND LIVES 33-41 (2008) (discussing use and misuse of statistical testing in different fields).

she gives a placebo to some subjects and the drug to be tested to others, the basic approach being to see whether the two groups react differently. Of course, the researcher recognizes that even if she does observe differences in the reactions of the two groups she cannot be certain that those differences were caused by the drug and not some other reason(s). For instance, the two populations might be dissimilar in ways of which she was unaware. If they were, then one or more of these unknown variables—and not the drug—might explain the differences observed. Even if the two groups were identical in every way, it might also just be a matter of chance that the reactions of the two groups were different.

Our drug researcher would like to be able to answer how unlikely it is that the differences she observes are the result of any of these other rival hypotheses, including the rival hypothesis chance. However, for reasons that are historical, complicated and not necessarily defensible, in biomedical research, as well as in the social sciences, the accepted practice is to begin by assessing the degree to which the researcher can be confident that the rival hypothesis chance is not the reason for the results.[44]

Statisticians use what is called, rather confusingly, null hypothesis statistical testing to gauge the probability that an association or difference between two variables would be found that is "as or more extreme than the one observed if the association or difference existed only by chance."[45] The p-value that is computed by statisticians is the numerical value given to that probability.[46] The acronym is confusing, however, because strictly speaking one cannot use null hypothesis statistical testing to measure the probability that the tested hypothesis of no effect (the most commonly used null hypothesis) is true. The eminent psychologist Jacob Cohen famously put it this way in his paper, *The Earth is Round (p <.05)*:

> What's wrong with NHST [null hypothesis statistical testing]? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is 'Given these data, what is the probability that H_0 is true?' But as most of us

---

[44] *See* Ronald P. Carver, *The Case Against Statistical Significance Testing*, 48 HARV. EDUC. REV. 378, 380, 382 (1978) ("Statistical testing sets up a straw man, the null hypothesis, and tries to knock him down. . . . When calculating the p values, we assume…that all the odds are in favor of chance causing the results.").

[45] Richard Lempert, *The Significance of Statistical Significance: Two Authors Restate an Incontrovertible Caution. Why a Book?*, 34 L. & SOC. INQUIRY 225, 232 (2009).

[46] *See generally* Carver, *supra* note 44 at 380 ("When the null hypothesis is used in research, the known variability in the sampled groups can be used to *estimate* the unknown variability in the assumed common populations. Using this estimate of the population variability and the known sample size, we can mathematically calculate how often we would expect to find mean differences—sampling errors—of any particular size. The calculations from a *t* test provide a p value. . . a number which tells us the *proportion* of the time that we can expect to find mean differences as large or larger than the particular sized difference we get when we are sampling from the same population assumed under the null hypothesis") (emphasis in original).

> know, what it tells us is 'Given that H_0 is true, what is the probability of these (or more extreme) data?'[47]

Cohen is reminding us that because the p-value says something only about the *data*, not about the *hypothesis* being tested, it can only denote the probability of effects recurring in future experiments; and because it is based on an initial assumption that the null hypothesis is true, it cannot—by definition—tell us the probability that the null hypothesis is actually correct or incorrect.[48] In statisticians' parlance, $P(D \mid H_0) \neq P(H_0 \mid D)$.

If our drug researcher were to come up with a p-value of .05 or smaller, what that tells her is that the probability is one in twenty that she would have observed effects in the size (or even larger effects) if the difference between the observed and treatment groups were the result of chance variations as they affected the outcomes in the experimental and control groups. A small p-value is some evidence that the observed effects were the result of a real association with the drug, and not random error, but is not the same as saying that her calculated p-value proves a causal relation between the drug and the effects she detected. As Richard Lempert has put it, "[r]ejecting a null hypothesis is not the same as proving a favored one."[49] There are still all of those other rival hypotheses to consider—which the p-value, calculated based only on an assumption that the null hypothesis of no effect is true, tells us nothing about. Correspondingly, when a finding is not statistically significant, that "does not mean that the null hypothesis is true" even though, as Michael Green has noted, this is a very common misconception.[50]

By failing to clarify the meaning of a finding of no statistical significance, the FJC study confused readers into thinking that it proved the Court's cases had no impact on dismissal practice. As previously stated, null hypothesis statistical testing can do no such thing. By clarifying what a p-value denotes and what it does not denote, the researchers could have aided understanding of how to interpret the study's results.

Beyond confusion over the meaning of a no-significance finding, a persistent error among researchers and readers alike has been to mistake statistical significance for practical or substantive importance. Especially among those untrained in statistical analysis, there is a tendency to assume that findings that are not statistically significant can be dismissed as untrue or not important.[51] A statistically significant result might be something that we care very little about—and it does not become substantively more significant as the

---

[47] Jacob Cohen, *The Earth is Round (p < .05)*, 49 AM. PSYCHOLOGIST 997, 997 (1994).

[48] *See id.* at 998-99*; see also* Carver, *supra* note 44 at 382) (explaining that "the *p* value is the probability of getting the research results when it is first assumed that it is actually true that chance caused the results. It is therefore impossible for the *p* value to be the probability that chance caused the mean difference between the two research groups since (a) the *p* value was calculated by assuming that the probability was 1.00 that chance did cause the mean difference, and (b) the *p* value is used to decide whether to accept or reject the idea that probability is 1.00 that chance caused the mean difference").

[49] Lempert, *supra* note 45, at 235.

[50] Michael D. Green, *Legal Theory: Expert Witnesses and Sufficiency in Toxic Substances Litigation: The Legacy of Agent Orange and Bendectin Litigation*, 86 NW. U. L. REV. 643, 682-83 (1992).

[51] *See* Kirk, *supra* note 17, at 214.

value of p drops.[52] Correspondingly, and of particular relevance with regard to the FJC study of *Twombly* and *Iqbal*, saying that a relationship between variables is not statistically significant does not mean that the observed effects are unimportant.

Frank Yates, one of the leading statisticians of the last century sharply made the point more than a half century ago that null hypothesis significance testing "has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, and too little to the estimates of the magnitude of the effects they are investigating."[53] In legal academia, many excellent sources explain that statistical significance does not say anything about the size or importance of the results obtained.[54] Even the U.S. Supreme Court has realized the distinction, albeit as a latecomer, reminding us that (at least in the context of Rule 10b-5 securities actions) courts should not confuse statistical significance for substantive importance.[55]

Despite all of the awareness that statistical significance is not the same as practical importance, the "cult of statistical significance" persists and appears very hard to correct, as Stephen Ziliak and Deirdre McCloskey have recently demonstrated.[56] By emphasizing only whether the differences they observed were statistically significant, the FJC researchers fell into a common trap that may have led at least some of its readers into confusing statistical for

---

[52] Cohen, *supra* note 47, at 1001; *see also* Alan G. Sawyer & J. Paul Peter, *The Significance of Statistical Significance Tests in Marketing Research*, 20 J. MARKETING RES. 122, 123 (1983), *available at* www.bauer.uh.edu/jhess/documents/Sawyer%20and%20peter.pdf (referring to "the practice of interpreting p-values as a measure of the degree of validity of research results, i.e., *p*-value such as $p < .0001$ is 'highly statistically significant' or 'highly significant' and therefore much more valid than a *p*-value of, say, .05," and noting that "such a practice is inappropriate").

[53] F. Yates, *The Influence of Statistical Methods for Research Workers On the Development of the Science of Statistics*, 46 J. AM. STAT. ASS'N 19, 33 (1951).

[54] *See, e.g.*, D.H. Kaye, *Is Proof of Statistical Significance Relevant?*, 61 WASH. L. REV. 1333, 1356-64 (1986); Epstein & King, *supra* note 42, at 49-55; Epstein, Martin & Schneider, *supra* note 42, at 1827-44.

[55] *Matrixx Initiatives Inc. v. Siracusano*, 131 S. Ct. 1309, 1319 (2011) ("A lack of statistically significant data does not mean that medical experts have no reliable basis for inferring a causal link between a drug and adverse events."); *id.* (noting that "medical professionals and researchers do not limit the data they consider to the results of randomized clinical trials or to statistically significant evidence"). As it turns out, in the same decision that recognized the distinction between statistical and practical significance, the Court cited *Twombly* and *Iqbal* as authority for upholding the sufficiency of the plaintiff's complaint. *See id.* at 1323 ("We believe that these allegations suffice to 'raise a reasonable expectation that discovery will reveal evidence' satisfying the materiality requirement, *Bell Atlantic Corp.* v. *Twombly*, and to 'allo[w] the court to draw the reasonable inference that the defendant is liable for the misconduct alleged," *Iqbal* . . . . Viewing the allegations of the complaint as a whole, the complaint alleges facts suggesting a significant risk to the commercial viability of Matrixx's leading product.") (citations omitted). It is hard to gauge which is more uncertain: that the lower courts will follow the Court's explicit criticisms in *Matrixx* of over-relying on statistical significance or its more opaque reference to *Twombly/Iqbal* that might or might not suggest an intended softening of the pleading sufficiency standard. For more on *Matrixx*, *see* D.H.. Kaye, *Trapped in the* Matrixx*: The U.S. Supreme Court and the Need for Statistical Significance*, 39 PROD. SAFETY & LIAB. REP. 1007 (2011) (analyzing Supreme Court's discussion of statistical significance testing in *Matrixx*).

[56] ZILIAK & MCCLOSKEY, *supra* note 43, at 55 and 238-44 (discussing and explaining continued influence of "cult of statistical significance").

substantive significance. Whether statistically significant or not, with recognition of the limited meaning of a no-significance finding, readers—and policymakers, in particular—could reasonably be concerned by the higher number and rate of orders granting dismissal since *Iqbal* both overall and in every case category examined.

B. Clarifying the Costs of Setting an Inappropriately High Threshold for Statistical Significance

Additionally, the researchers failed to make clear that in employing the statistical test that they used they were measuring only the likelihood of a false-positive error—that is, thinking the Court's decisions were responsible for the higher dismissal rates when, in fact, they were not.

As noted above, the purpose of doing null hypothesis significance testing is to gauge the probability that an association or difference between two variables would be found that is as extreme or more extreme than the one observed if the null hypothesis were true (i.e., that no real association or difference exists). A statistical level of significance set at .05 means that the researcher is essentially measuring only one type of error: the probability that one would think there is an effect when, in fact, there is not.[57] Beyond the possibility of making a false-positive mistake, there is, of course, another kind of error: the possibility that one would not think there is an effect when, in fact, there is one (a Type II or false-negative error). Both kinds of errors should matter. We would not want our drug researcher to conclude erroneously that the drug she investigated had positive effects that it did not actually have. We also should be concerned that the researcher will erroneously think the drug had no effect when, in truth, its effects were highly beneficial.

With regard to epidemiological studies, Michael Green discusses the inverse relationship between the two types of error. Green notes that the more one worries about making a false positive error, the greater the chances of missing effects that are really there.

> Requiring that a study finding a positive association be statistically significant before it is treated as probative for purposes of proving an effect will almost always produce more error of the false negative variety (favoring defendants) than the false positives (favoring plaintiffs) it avoids. This results because statistical significance demands that an effect must be a result that would occur through random chance less than five percent of the time. All outcomes, no matter how high the magnitude of the effect, are rejected unless there is a probability of five percent or less that the effect found is due to chance. But, this minimization of false positives comes at a cost—rejecting the existence of an effect when one truly exists—a false negative, which is beta error.[58]

---

[57] Green, *supra* note 50, at 682.

[58] *Id.* at 687.

Green reminds us that the tension between Type I and Type II errors demands that in setting the appropriate significance level one must consider the study's purposes and the relative costs of both types of errors. The FJC researchers were right to try to minimize the likelihood of a false positive error. However, in addition, they should have made clear to readers that a false negative error—here, the possibility of thinking that *Twombly* and *Iqbal* were not responsible for the increases in the dismissal grant rate when, in fact, they were—also matters. Indeed, one might reasonably decide, given the stakes involved in terms of judicial access, that a false negative error matters more.

Readers—and especially the policy makers who commissioned this study—would have been far better served if the researchers had made clear the danger of ignoring the post-*Iqbal* differences that were found in dismissal practices and outcomes merely because they were not statistically significant at the .05 level. To readers untrained in statistical testing, this means they should have underlined the point plainly that using the conventionally high bar of .05 for statistical significance carries the real danger that actual causal effects will be missed. Even more importantly, the researchers should have tried to determine the probability of false-negative error, and not merely limited their analysis to Type I error. This could have been done by measuring the "power" of the study—an analysis that allows researchers to determine "the likelihood that a study will find a statistically significant hypothesized effect if a real association exists."[59] By considering the study's power, the researchers could have made a better estimate of the appropriate significance level to use by taking into consideration the costs of both types of errors. Their estimates might have been subsequently questioned, but at least the discussion would have been transparent and, thereby, could have helped focus policymakers' attention where it should be: not on summary declarations of statistical significance but on how to evaluate the study's empirical findings in light of the need to balance both kinds of potential errors. This work is not easy, but it is the kind of effort that is demanded both of empiricists and policymakers.

C. Raising Plausible Reasons Why the Observed Differences Might Not Have Been Statistically Significant

I have argued so far that to clarify to readers the limits of the study's results, the researchers should have discussed what a finding of no statistical significance means, as well as what the costs are of setting an inappropriately high threshold for statistical significance. Finally, I argue here that the researchers should also have made clear the limits of their study's results by noting that there are numerous reasons why the observed differences might not have reached the predetermined level of statistical significance they selected.

1. The Study Should Have Emphasized that Some of the Assumptions Underlying their Empirical Models May Not Have Been Correct

One of the reasons the authors could have emphasized to readers why the

---

[59] *Id.* at 685.

differences they observed might not have reached statistical significance at the .05 level is that the empirical models used may not have been correctly constructed. One of the assumptions made by the FJC researchers in constructing the model they used to calculate the statistical significance of the findings was that it was appropriate to compare the 2010 dismissal rates with a pre-*Twombly* "baseline" of dismissal rates using a single category of cases (tort cases) from the mean dismissal rate of three districts (the District of Rhode Island, the Eastern District of Michigan and the District of Maryland). Although the effort may have been well intended, it is not clear that it was right to regress from this constructed pre-*Twombly* "baseline." As Table A-2 in the appendices to the FJC study shows, controlling for different dismissal rates in different courts, the grant rates in every other case category were lower in 2006 than the constructed baseline, raising concern that the model's comparison of the dismissal rate understated the increase in the later period.

A still further and even more potentially problematic assumption is that the variables selected truly were independent of *Twombly* and *Iqbal*'s effects. The regressions the FJC researchers ran controlled for three variables: court, case type and whether the order responded to an amended complaint. The results of their models reveal valuable information, however, only if these variables really are independent of *Twombly* and *Iqbal*'s effects; yet it is not clear that they all are truly independent. For instance, why should we assume the district court is entirely independent of *Twombly* and *Iqbal*'s effects? It is even less clear why it is appropriate to isolate out whether the court's order was in response to a complaint that had been amended. As noted earlier, an increase in the grant rate may be alarming even when leave to amend has been given, especially when the FJC's own data show that the movant's success rate goes up significantly after the plaintiff has had an opportunity to amend. Further still, the researchers chose to isolate all of the variables when it might have been more appropriate to take only a single variable—say, case type—and stratify within each case type other variables, such as the different courts. Changing the model in these ways would have produced different test results, and general readers should have been made aware of the model's susceptibility to modification to help increase understanding of the study's results.

2. The Observed Differences Might Not Have Been Statistically Significant Because of the Size of the Sample Studied

Readers would also have been aided by an understanding that the observed differences might not have been statistically significant because of the size of the sample studied. Statistical significance depends in part on sample size, as well as on the size of the effect, the variability in the population, and other factors.[60] Very often, the larger the sample, the more likely it is that the researcher will be able to observe an effect and, by extension, the smaller the sample, the less likely it is that the effect will be detected. As Ronald Carver has explained, "statistically nonsignificant results are conventionally interpreted as providing no support for the research hypothesis even when the actual results support it. When a small sample is used, large differences in the

[60] Carver, *supra* note 44, at 386.

results can more often occur by chance and therefore provide no statistically significant evidence in support of the research hypothesis."[61] Richard Lempert has made the same point for a law journal audience:

> An even greater threat to science-based understandings is the problem of low power. Particularly when samples are small, even strong relationships may not be statistically significant. This may lead researchers to report [a] finding of no relationship in the data when a relationship not only exists but is substantively important.[62]

The culprit of small sample size is part of the story with the statistical testing that was done as to the FJC's findings in its *Iqbal* study, as a quick scan of the individual case categories readily reveals. For example, consider the Torts cases. As shown in Table 4 of the study, the total number of observed orders since *Twombly* and after *Iqbal* was, respectively, 15 and 32. Employment Discrimination cases were even smaller (just 17 orders pre-*Twombly* and 28 orders post-*Iqbal*).

Table 4: Outcome of Motions to Dismiss for Failure to State a Claim

	Action on Motion	2006	No. of Orders	2010	No. of Orders	Difference
Torts	Denied	30.0%	(21)	28.2%	(31)	
	Granted All or Some Relief	70.0%	(49)	71.8%	(79)	+1.8%
	With Amendment	21.4%	(15)	29.1%	(32)	+7.7%
	Without Amendment	48.6%	(34)	42.7%	(47)	-5.9%
Civil Rights	Denied	27.9%	(51)	22.0%	(51)	
	Granted All or Some Relief	70.3%	(121)	78.0%	(181)	+7.7%
	With Amendment	21.1%	(38)	32.8%	(76)	+11.7%
	Without Amendment	48.3%	(83)	45.3%	(105)	-3.0%
Employment Discrimination	Denied	32.6%	(31)	29.4%	(35)	
	Granted All or Some Relief	67.4%	(64)	70.6%	(84)	+3.2%
	With Amendment	17.9%	(17)	23.5%	(28)	+5.6%
	Without Amendment	49.5%	(47)	47.1%	(56)	-2.4%

* $p \leq 0.01$, relative to the likelihood that the motion will be denied.
† $p \leq 0.05$, relative to the likelihood that the motion will be granted without leave to amend.

Table 4 is reprinted from FJC Study at 14.

With sample sizes this small, to say that the differences detected were not statistically significant is not saying very much at all. It does not mean that the

[61] *Id.*
[62] Lempert, *supra* note 45, at 236.

Court's decisions are not responsible for the higher number of orders and higher grant rate in both of these categories (as we will discuss below). Nor does it say much about the magnitude of the effects observed.

In addition to the problem of small sample size, the size of the effect also matters in measuring statistical significance, as noted above. Effect size poses a unique challenge to researchers studying Rule 12(b)(6) dismissal orders, as Kevin Clermont and Stephen Yeazell have pointed out. Clermont and Yeazell note that not all Rule 12(b)(6) motions are alike and only some will be pure *Twombly/Iqbal* motions that challenge the factual sufficiency of allegations.[63] They submit that because pure *Twombly/Iqbal* motions will constitute only a percentage of all motions to dismiss for failure to state a claim, the effects of the cases will be masked by the non-*Twombly/Iqbal* motions. In another paper published before the FJC study, Clermont further elaborates on the point:

> [I]f one were to compile all dismissal decisions, the effects of *Twombly-Iqbal* would be hard to measure because these precedents apply to only a restricted subset of motions to dismiss (and result in final dismissal for a smaller subset). That is, *Twombly-Iqbal* will have its bite only in cases in which the plaintiff cannot plead more detail and the plaintiff nevertheless sues without the detail. The other cases will overwhelm and mask the subsets. In other words, the numbers of motions and dismissals might be high enough to conceal any effect of the new regime.[64]

When the FJC reported that the changes in dismissal grant rate were not statistically significant, it perhaps should have come as no surprise. Because pure *Twombly/Iqbal* motions are only one kind of Rule 12(b)(6) motion, Clermont predicted that it would be very hard to find statistically significant evidence of *Twombly* and *Iqbal*'s effects through gross quantitative efforts like those undertaken by the FJC study.[65]

3. The Study Should Have Reported the Actual Test Results and, Separately, Made Clear that at a Lower Level Many of the Effects Observed Would Have Been Statistically Significant

The study should also have reported the actual test results and, separately, made clear that at a lower level many of the effects observed would have been statistically significant. The problem begins with a lack of transparency insofar as the researchers did not report their actual test results. The common

---

[63] Clermont & Yeazell, *supra* note 5, at 839 n.66.

[64] Clermont, *supra* note 9, at 1366 n.140.

[65] *Id.* (observing that "when I contemplate the possibility of a relatively noninflated numerator and an inflated denominator in the dismissal success rate, combined with the inevitable case-selection effect, I am left wondering whether any study looking at the numbers of motions and dismissals really could result in anything other than a showing of little impact").

practice in scientific journals is to report actual p-values, rather than merely reporting results as significant or not.[66]

To illustrate how summary declarations of significance or nonsignificance abet misunderstanding, consider how the findings were reported for Civil Rights cases. The researchers reported that the grant rate in Civil Rights cases increased after *Iqbal* (as the excerpt from Table 4 shows, it went from 70.3% to 78%), but that this increase was not statistically significant at the .05 level:

Table 4: Outcome of Motions to Dismiss for Failure to State a Claim

	Action on Motion	2006	No. of Orders	2010	No. of Orders	Difference
Civil	Denied	27.9%	(51)	22.0%	(51)	
Rights	Granted All or Some Relief	70.3%	(121)	78.0%	(181)	+7.7%
	With Amendment	21.1%	(38)	32.8%	(76)	+11.7%
	Without Amendment	48.3%	(83)	45.3%	(105)	-3.0%

* $p \leq 0.01$, relative to the likelihood that the motion will be denied.
† $p \leq 0.05$, relative to the likelihood that the motion will be granted without leave to amend.
Table 4 is reprinted from FJC Study at 14.

Though the p-value was not reported, the researchers subsequently confirmed that they calculated it as .08, a result that is not significant at the .05 level but would have been significant at .10.[67] This, in turn, raises the point that it would have aided understanding among general readers for the researchers to have acknowledged that any particular threshold of significance level is necessarily arbitrary and that at a lower level many of the observed differences would have been statistically significant.[68] Even Ronald Fisher, who more than anyone influenced science's adoption of .05 as the conventional level of statistical significance, acknowledged the arbitrariness of the cutoff and urged researchers to report the exact figures, rather than relying on summary declarations of significance.[69]

4. The Study Should Have Discussed the Difference Between One-Tailed v. Two-Tailed Tests and Pointed Out that a Plausible Argument Can be Made in this Context for Using the One-Tailed Test

Finally, the researchers also could have acknowledged that although they used an accepted test, known as a two-tailed test, to calculate the p-values, a justifiable argument can be made in this context for using a different test

---

[66] Kaye, *supra* note 48, at 1344.

[67] Email correspondence from Joe Cecil to Lonny Hoffman, August 2, 2011 (copy on file with author).

[68] *See generally* Thomas W. Nix & J. Jackson Barnette, *The Data Analysis Dilemma: Ban or Abandon. A Review of Null Hypothesis Significance Testing* 5 RESEARCH IN THE SCHOOLS, Fall 1998, at 3; Dominic Beaulieu-Prevost, *Confidence Intervals: From Tests of Statistical Significance to Confidence Intervals, Range Hypotheses and Substantial Effects*, 2 TUTOR. IN QUANT. METHODS FOR PSYCHOLOGY 11 (2006).

[69] Kaye, *supra* note 48, at 1343-44 (citing and quoting Fisher's work).

(known as a one-tailed test) for evaluating statistical significance. Even though the two-tailed test usually will be the more appropriate test to use, reasonable arguments can be made in favor of the one-tailed test in this context, since the effects of the Court's cases are likely to be unidirectional (that is, when it is difficult to believe that a stricter pleading test would lead to fewer dismissals).[70]

Had the one-tailed test been used, it would have yielded a p-value of .0391 for the increase in the rate at which motions to dismiss in Civil Rights cases were granted, which would have made the increase statically significant even at the conventional five-percent level. A one-tailed test also would have resulted in a lower p-value than the FJC researchers reported for cases in the "Other" category, the largest case category. Cases in the "Other" category included antitrust, RICO, ERISA, copyright, patent, environmental, other statutory actions, and a number of other case types. The one-tailed p-value for the grant rate increase for Other is .0539. Not at the .05 level, but darn close. How close? Put it this way: if the researchers had miscoded even a single case, making the grant rate 197 instead of 196 (as reported), then p=.0448 and it is significant at .05 (a graphic illustration of how the FJC's model was quite fragile to small changes in the data). The point is not that the one-tailed test was necessarily the more appropriate test to employ. It is, instead, that the researchers should have made clear to readers that if they had employed this alternative test (for which, as just noted, reasonable arguments can be made for its use) most of the grant-rate increases would have been statistically significant (or just nearly so) even at the conventional .05 level.

* * *

Summing up, I have argued that the researchers should have made clear the limits of their study's results by clarifying what a finding of no statistical significance means; by explaining the costs of setting an inappropriately high threshold for statistical significance; and by noting numerous plausible reasons why the differences they observed might not have reached the predetermined level of statistical significance they selected. If they had done so, readers might have been less likely to assume that the study demonstrated that *Twombly* and *Iqbal* were not responsible for the higher number and rate of dismissals and to overlook the considerable effects the researchers observed in the post-*Iqbal* period.

IV. WHAT THE RESEARCHERS COULD NOT DETECT

I have argued that the researchers failed to help readers understand what it

---

[70] DAVID W. STOCKBURGER, INTRODUCTORY STATISTICS: CONCEPTS, MODELS, AND APPLICATIONS 193 (1996) (noting that the "one-tailed t-test is performed if the results are interesting only if they turn out in a particular direction"); *see also* Alan O. Sykes, *An Introduction to Regression Analysis*, in CHICAGO LECTURES IN LAW & ECONOMICS 1, 22 (Eric A. Posner ed., 2000) (using gender discrimination to illustrate use of one- and two-tailed tests and noting, inter alia, that "we may regard the two-tailed test as inappropriate for the coefficient of the gender dummy because we find the possibility of discrimination against men to be implausible").

.

means to say that the observed differences in dismissal outcomes after *Iqbal* were not statistically significant. In addition, it is equally vital to keep in mind what the researchers could not detect. To their credit, at various times in their deliberations, rulemakers have recognized the limits of empirical study of *Twombly* and *Iqbal*'s effects.[71] The FJC researchers themselves likely understood the limits of their investigation; none of their findings are presented as policy recommendations. Nevertheless, it may be that the limits of empirical research into *Twombly* and *Iqbal*'s effects are too easily forgotten when a comprehensive study by distinguished researchers is presented in such a way that it unintentionally suggests the cases are not having the kind of serious, systematic changes in Rule 12(b)(6) activity that had been anticipated.

A. The FJC Study Could Not Measure How Many Prospective Claimants Were Deterred by *Twombly* and *Iqbal* from Seeking Relief

One difficulty in assessing *Twombly* and *Iqbal*'s effects is that a study comparing pre-*Twombly* and post-*Iqbal* filing rates and movant success rates does not tell us how many prospective claimants were deterred from seeking legal relief because of the Court's more exacting pleading standard. Indeed, it is not clear how any empirical study could measure the deterrent effect of the Court's decisions. One suggestion that has been offered is that we might look at the total number of lawsuits filed. That approach, however, does not seem likely to shed much light on the deterrence problem since so many different variables influence the case filing rate.[72]

Some empirical work that has been conducted involving securities cases suggests that the Private Securities Litigation Reform Act's heightened pleading requirement has resulted in some meritorious cases not being filed. In their July 2007 study of securities class actions involving allegations of secondary market fraud, Stephen Choi, Karen Nelson, and A.C. Pritchard found that the PSLRA's heightened pleading standard has had "screening effects," as they call it.[73] With respect to suits that would have settled for non-nuisance value (their shorthand for a meritorious case) before the PSLRA, the authors found that:

---

[71] *See, e.g.*, Report to Standing Committee from Civil Rules Committee, in Agenda Materials, June 2011 Meeting of Committee on Rules of Practice and Procedure, at 53 (noting that "[o]ther questions elude the capacities of even the most careful docket studies. It is not possible to identify cases that would have been filed under earlier understandings of pleading standards but were not filed for fear of heightened pleading standards. . . . It is not possible to determine whether cases were dismissed for want of pleading facts that could be known only by discovering information available only by discovery from the defendant. It would be difficult to assess the quality of the differences between initially unsuccessful complaints and successful amended complaints, or to measure the advantages of an amended complaint in working toward ultimate resolution. And it is similarly difficult to distinguish pleadings that fail for want of factual sufficiency alone and those that fail in whole or in part for advancing an untenable legal theory").

[72] For an effort at trying to measure, *inter alia*, *Twombly*'s impact on the case filing rate, see William Hubbard, *The Problem of Measuring Legal Change, with Application to* Bell Atlantic. v. Twombly, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1883831.

[73] Stephen J. Choi et al., *The Screening Effect of the Private Securities Litigation Reform Act*, 6 J. EMP. LEGAL STUD. 35 (2009).

> [A] substantial percentage of suits that would have resulted in a nonnuisance settlement prior to the PSLRA would not have been filed after Congress adopted the PSLRA. The screening effect diminishes, however, if we consider cases with "hard evidence" of securities fraud – a restatement of earnings or revenues or an investigation by the SEC – or abnormal insider trading.[74]

In other words, they ascertained a filing deterrence effect as a result of the PSLRA heightened pleading requirement that was most pronounced in cases in which access to hard evidence of wrongdoing was not as readily accessible to the plaintiff. In sum, as Choi, Nelson, and Pritchard observe, "[o]verall, our findings do not show that Congress's efforts to discourage frivolous litigation have succeeded; indeed, we find stronger evidence that the PSLRA has succeeded in discouraging securities fraud class actions that would likely have been deemed meritorious prior to the PSLRA."[75]

Anticipating the problem, Arthur Miller underlined the danger concisely, keying in on concern about the kind of cases that might be deterred by *Twombly* and *Iqbal*:

> [T]he plausibility-pleading standard risks increased difficulty for many prospective claimants—some with claims that may well have merit and involve important public policies—to survive a Rule 12(b)(6) motion. In an unknowable number of instances, the increased risk of dismissal and the resources needed to defend against it may deter the institution of a potentially meritorious case. . . . This is especially worrisome in cases involving important issues—such as constitutional values and the private enforcement of federally and state-created rights—and the concomitant shift in the allocation of the litigation-resource burden from defendants to plaintiffs these two decisions produce. The result is likely to operate in derogation of effectuating rights and policy norms established by Congress and state legislatures.[76]

The FJC study was not designed to determine whether a similar deterrence effect was occurring among prospective claimants as a result of the Court's decisions. The study, thus, cannot tell us whether *Twombly* and *Iqbal* are causing some who have been wronged not to file meritorious claims out of a

---

[74] *Id.* at 37.

[75] *Id.* at 65. Even among the group not deterred from filing suit, other adjustments to the new pleading regime may have to be made that would not be observable. For instance, more factual detail may be going into complaints, presumably causing at least some claimants to incur additional costs to gather the necessary additional detail perceived to be necessary to meet the Court's new pleading requirements. *See* 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 (2010) ("Plaintiffs are required to produce a considerable degree of factual detail at the very beginning of the lawsuit before they have been able to conduct any discovery.").

[76] Miller, *supra* note 7, at 47, 77 (footnotes omitted).

concern they would not be able to meet the general pleading requirement of Rule 8.[77]

B. The FJC Study Could Not Measure How Often Meritorious Cases Have Been Dismissed Under the *Twombly*/*Iqbal* Test

Empirical study of Rule 12(b)(6) activity also cannot tell us how often cases are being dismissed at the pleading stage that, if allowed to proceed to discovery, would have resulted in production of evidence to support a meritorious claim. This possibility could arise any time that the plaintiff lacks access to proof of wrongdoing that is solely in the defendant's possession. I drew attention to the problem of information asymmetry after *Twombly* was announced;[78] and although it remains difficult to determine how often this problem arises, it is one of the key policy questions that rulemakers must address. Discovery rule reform proposals currently being considered by rulemakers could help ameliorate the information asymmetry problem, but they are necessarily only a partial and inadequate remedy for all of the concerns that *Twombly* and *Iqbal* trigger when imbalances in critical information exist. For now, the key point to be made is a study comparing grant rates before *Twombly* and after *Iqbal* is unable to tell us how many meritorious cases have been dismissed under the *Twombly*/*Iqbal* standard. That information, critical to know before we can make any assessment of the Court's new doctrine, is undetectable by the empirical methods used in the FJC's study.

C. The FJC Study Could Not Detect Whether *Twombly* and *Iqbal* Have Significantly Increased Dismissals of Complaints for Being Factually Insufficient

Finally, the FJC's *Iqbal* study also does not tell us anything about the kinds of motions being filed and granted. More precisely, the FJC study cannot tell us whether the Court's decisions have significantly increased dismissals of complaints on the ground that they are factually insufficient. Even after *Iqbal*, courts are being asked to decide motions to dismiss on grounds that would have justified dismissal even before *Twombly* (such as a legal-sufficiency challenge). At the same time, even before *Twombly* and *Iqbal*, pure notice pleading was probably not practiced, at least not routinely, in the lower courts. That is, even before *Twombly*, defendants were seeking dismissals—and judges were granting them—on factual insufficiency grounds at least akin to the factual sufficiency review the Court authorized in *Twombly* and *Iqbal*. Given these two realities, to really evaluate *Twombly* and *Iqbal*'s effects what we would need to know is how often (i) defendants are filing the kinds of motions to dismiss that they would not have filed before *Twombly* and

---

[77] Robert G. Bone, *Plausibility Pleading Revisited and Revised: A Comment on* Ashcroft v. Iqbal, 85 NOTRE DAME L. REV. 849, 852 (2010) ("*Iqbal* applies a thick screening model that aims to screen weak as well as meritless suits, whereas *Twombly* applies a thin screening model that aims to screen only truly meritless suits. The thick screening model is highly problematic on policy grounds.").

[78] Hoffman, *supra* note 4, at 1260-64.

Iqbal and (ii) courts are granting those motions when they would not have done so before. Counting noses does not get at any of these deeper evaluative needs.

For instance, assume that there has been a ten-fold increase in factual-sufficiency challenges but, at the same time, a corresponding decrease in legal-sufficiency challenges. A study that compares the total volume of Rule 12(b)(6) activity before *Twombly* and after *Iqbal* could find total activity levels unchanged and entirely miss those dramatic changes actually taking place. If *Twombly* and *Iqbal* have increased the number of dismissals sought and/or granted because a claim was deemed to be factually insufficient, that would constitute a significant change in dismissal practice. Many academic commentators have argued that the central infirmity with the Court's decisions is that they empower judges to decide whether a case has merit at the pleading stage, confusing pleading sufficiency with the kind of evidentiary evaluation undertaken at summary judgment, routinely after discovery.[79] The FJC study cannot tell us whether the Court's decisions have transformed the nature of Rule 12(b)(6) challenges in this manner. As it turns out, preliminary results from a recently completed separate study seem to indicate that factual-insufficiency dismissal rates are much higher after *Iqbal* and, separately, that legal-insufficiency challenges are down.[80]

V. INCLUSIVENESS CONCERNS: DID THE FJC CAPTURE ALL OF THE RELEVANT ACTIVITY?

The previous part showed that there are important effects the cases may be having that the FJC researchers would not have been able to observe. I argued, therefore, that the researchers should have been clearer in reporting their findings to describe the limits of their empirical investigation expressly. In this final part, I set all prior criticisms aside. Taking the study on its own terms, I explore the possibility that the data the FJC researchers gathered may be incomplete. Consequently, I argue that there are reasons to be concerned that the study may be providing us an incomplete picture of actual Rule 12(b)(6) activity.

A. Discrepancies Between the Filing Rate Found in the 2011 Study and Two Prior Studies of Rule 12(b)(6)

We saw earlier that the FJC found that motions to dismiss for failure to state a claim were filed in 4% of all cases in the pre-*Twombly* period running from October 2005 through June 2006. This may be one of the most startling findings in the study. The 4% filing rate is significantly lower than the rate found by two earlier studies of Rule 12(b)(6) going back several decades, both also conducted by the FJC. The first study found between a 15-18.7% filing rate for cases terminated in 1975.[81] The second study, completed in 1989,

---

[79] *See supra* note 9.

[80] SCOTT DODSON, SLAMMING THE FEDERAL COURTHOUSE DOORS: NEW PLEADING IN THE TWENTY-FIRST CENTURY (forthcoming 2013) (copy of draft on file with author).

[81] PAUL R.J. CONNOLLY & PATRICIA A. LOMBARD, JUDICIAL CONTROLS AND THE CIVIL LITIGATIVE PROCESS: MOTIONS (Fed. Jud. Ctr. 1980) (18.7% figure based on 582 motions

observed that Rule 12(b)(6) motions were filed in 13% of all civil actions.[82] The 2011 study cited these prior studies, noting the discrepancies, but did not address them further.[83] Why would the filing rate have fallen so dramatically (a decline of approximately roughly 70%) from the 1980s to 2005-06? The explanation for the dramatic decline from one period to the other (keeping in mind that both periods, of course, were pre-*Twombly*) is not immediately apparent. Indeed, the decline in the filing rate is particularly puzzling since the prior evidence indicates that the Rule 12(b)(6) filing rate has held very steady over the several decades in which such data have been gathered.[84]

The FJC's finding that the post-*Iqbal* filing rate was 6.2% across the 23 districts in all case types is equally surprising. Only a year after the *Twombly* decision was announced, it was noted that the case had already been cited with great frequency,[85] suggesting that defendants were "now more regularly urging judges to intercept complaints at the pleading stage."[86] Several other commentators similarly predicted that after *Iqbal* defendants would be more routinely challenging the sufficiency of the plaintiff's pleadings.[87] In this same connection, it is perhaps notable that the 6.2% post-*Iqbal* filing rate the FJC found seems at odds with survey results of lawyers with the National Employment Lawyers Association (NELA), in which nearly three quarters reported that they had responded to motions to dismiss they believed would not have been brought before *Twombly*.[88]

B. Some Possible Explanations for the Discrepancies

---

filed out of 3,114 total cases in study). See *id.* at 70-71 (Tables 19 and 20). Note, however, that Connolly and Lombard defined "motions" to include court-initiated orders as well as party-initiated motions. *See id.* at 70 (footnote under Table 19). A later Federal Judicial Center study described the earlier Connolly and Lombard study as having found that motions to dismiss for failure to state a claim were filed in 15% of the cases. *See* Willging, *supra* note 16 at 3 (referring to the Connolly & Lombard study and characterizing it as finding Rule 12(b)(6) motions filed in "approximately 15% of the cases"). The discrepancy may be explained by a possible later effort by Willging to separate duplicate "motions" in light of the earlier researchers' decision to aggregate court-initiated orders and party-initiated motions. *Compare id.* at 5 n.10 (using figure of 582 motions and 3,114 total cases from Tables 19 and 20 of Connolly and Lombard study and referring to a 19% filing rate) with *id.* at 6 (Table 1) (reporting a separate figure from the 1980 study of 462 "Cases with Rule 12(b)(6) Motions" and providing 15% as the percentage of the sample in which such motions were filed).

[82] Willging, *supra* note 16, at 8.

[83] CECIL *et. al*, *supra* note 11, at 10-11 n.21.

[84] *See* WILLGING, *supra* note 16, at 5 (noting that "empirical data show a modern, consistent use of such motions to dispose of cases and claims" and summarizing prior research on filing rates of motions to dismiss for failure to state a claim going back to 1975).

[85] *See* Hoffman, *supra* note 4, at 1222-23.

[86] *Id.* at 1223.

[87] *See, e.g.*, Clermont & Yeazell, *supra* note 5, at 840 (observing that after *Iqbal* "any defendant's lawyer, faced with a complaint employing the minimalist pleading urged by Rule 8's wording and the appended Forms' content, commits legal malpractice if he or she fails to move to dismiss with liberal citations to *Twombly* and *Iqbal*") (footnote omitted).

[88] EMERY G. LEE III & THOMAS E. WILLGING, ATTORNEY SATISFACTION WITH THE FEDERAL RULES OF CIVIL PROCEDURE, REPORT TO THE JUDICIAL CONFERENCE ADVISORY COMMITTEE ON CIVIL RULES 12 (2010), available at http://www.fjc.gov/public/pdf.nsf/lookup/costciv2.pdf/$file/costciv2.pdf.

1. The 90-Day Cutoff

One explanation for the disparities found in the rate at which motions to dismiss for failure to state a claim were filed might be that in 2011 the FJC looked only at motions filed within ninety days of the case being brought. By contrast, the two older FJC studies catalogued motion to dismiss activity over the life of the cases examined. The researchers acknowledged that they may have missed some motion activity because of the cutoff date but were assuaged by a separate finding that the average time between case filing and filing of the first motion to dismiss was forty days—a figure that was basically the same both before *Twombly* and after *Iqbal*.[89] However, because they were looking only at a pool of motions filed within the first ninety days of a case's commencement, this finding means only that one who moves for dismissal within the first ninety days of a case will do so around the fortieth day. While it is often strategic for a defendant to seek dismissal early in the case, for many reasons that might not happen within the first ninety days. A defendant might not be served promptly after filing. A defendant might elect to waive service of process, thereby extending the time to file her answer up to sixty days. The parties might agree to extend answer and motion deadlines. A defendant might also move to dismiss allegations in a complaint that has been amended more than ninety days after the case was initially filed. Perhaps a defendant might seek dismissal—which she may do at any point in the case—if there has been a favorable change in the law. For any of these reasons, a substantial number of motions may have been brought more than three months after initial case filing.

The missing activity would be significant by itself, but the even more concerning question is whether the 90-day cutoff may have biased the results in one direction. There is reason to think it could have. After *Twombly* and *Iqbal*, many defendants might have concluded that the Court's decisions provided them an opportunity to seek dismissal that they previously did not have. As previously noted, several commentators predicted that that is exactly what defendants would conclude after *Iqbal*.[90] It is possible, therefore, that more defendants may have been led since *Iqbal* to seek dismissal of claims brought more than three months earlier, as compared to defendants who, before *Twombly*, had no similar incentive to seek dismissal if they had not already done so in the first ninety days.

2. Exclusion of Prisoner and *Pro Se* Cases

That the *Iqbal* study excluded prisoner and *pro se* cases seems a second likely explanation for at least some of the discrepancy in the filing rate found between this and the earlier studies. The key exclusion seems to have been prisoner and *pro se* cases.[91] Both the 1980 and 1989 FJC studies included

---

[89] CECIL, ET AL. *supra* note 11, at 8 n.13.

[90] *See supra* text accompanying notes 86-88.

[91] Most prisoner cases are *pro se*. *See* ROGER A. HANSON & HENRY W.K. DALEY, CHALLENGING THE CONDITIONS OF PRISONS AND JAILS: A REPORT ON SECTION 1983 LITIGATION 21 (1994) (noting that nearly all § 1983 suits brought by prisoners are *pro se* but also noting that many *pro se* cases are not brought by prisoners).

them. While these cases are a relatively small percentage of the entire civil docket,[92] if prisoner cases have a higher incidence of Rule 12(b)(6) activity than other civil cases then it seems likely that the exclusion of these cases, as well as the non-prisoner *pro se* cases, may explain some of the discrepancies in the filing rate between the 2011 study and the earlier FJC studies. Some prior research indicates that prisoner petitions, at least during the 1980s, had an above-average likelihood of involving Rule 12(b)(6) activity.[93]

However, if the exclusion of prisoner cases provides part of the explanation for the significantly lower pre-*Twombly* filing rate the FJC's *Iqbal* study found, as compared with the earlier studies, it cannot explain all of the differences. There are discrepancies not only in the overall filing rate, but also with regard to every case category studied.[94] For instance, Willging's 1989 study found a filing rate of 9% for employment discrimination cases, as compared with only 6.9% in the 2011 study. Additionally, Willging observed motions to dismiss for failure to state a claim in 14% of other civil rights (non-prisoner) cases, as compared with just a 9.7% filing rate found by the 2011 study. The consistently higher rates found by the 1989 FJC study for all case categories strongly suggests that the exclusion of prisoner cases does not explain all of the disparities in observed filing rates between the studies.

3. Other Possible Explanations: Coding Errors and Search Term Limitations

If neither the 90-day window nor the exclusion of prisoner cases explains all of the discrepancy in filing rates, what other possible explanations exist? Two other factors may have led the 2011 researchers to miss some Rule 12(b)(6) activity. The first has to do with the study's reliance on the CM/ECF coding by the clerks of potentially relevant motions. Only motions coded by the clerk under the event subcategory code "motion to dismiss" made it into the filing rate cohort that was collected. If a court clerk did not code a motion correctly, it would not have been included in the dataset.[95] While it is not possible to know how often miscodings may occur, a study that relies on the CM/ECF coding is susceptible to these sorts of errors. And it is worth noting that the miscoding problem does not go in both directions. That is, any coding

---

[92] For instance, from September 30, 2009 through September 30, 2010 there were approximately 283,000 private cases filed, of which approximately 25,000 were prisoner petitions regarding conditions and other civil rights claims (which are the kinds of petitions that can trigger Rule 12(b)(6) activity). *See* Administrative Office of the U.S. Courts, Statistical Tables for the Federal Judiciary, December 31, 2010, Table C-2. Habeas petitions made up another 20,000 cases, but Rule 12(b)(6) motions are rarely brought in habeas cases. For a rare exception see *Hopkins v. Grondolsky*, 759 F. Supp. 2d 97 (D. Mass. 2010).

[93] *See* WILLGING, *supra* note 16, at 7.

[94] The 1980 study by Lombard and Connolly did not break out Rule 12(b)(6) activity by case type, so the discrepancies noted in the text regarding case type are only between Willging's 1989 study and FJC's 2011 study.

[95] This might happen for several reasons (*e.g.*, a motion asking for relief on multiple grounds might have been coded for the relief first sought; a motion for dismissal might have been brought as part of the defendant's answer and so might have been coded only as an answer; a motion to dismiss might have been misnamed by the movant; or the clerk might have coded the dismissal motion mistakenly).

errors that led to wrongful inclusions either would have been filtered out by the FJC's subsequent electronic filtering or discarded from the sample manually by the researchers when they looked at the related orders and discovered them not to concern a motion to dismiss for failure to state a claim. In other words, coding error problems in this context are unidirectional. The miscoding of a motion to transfer venue under the event code "motion to dismiss" would not have affected the study's findings; but the miscoding of a Rule 12(b)(6) motion under the event code "motion to transfer venue" would.

A second explanation for some of the discrepancies in the filing rate between the 2011 study and the prior studies may be the search terminology used by the FJC researchers to cull Rule 12(b)(6) motions from the undifferentiated larger pool of "motions to dismiss." As noted above, the FJC drew its initial cohort of filings from all motions coded by the district clerks under the event subcategory "motion to dismiss." Because this general code is inclusive of motions seeking dismissal on any basis, it was necessary to identify within the cohort only those that sought dismissal at the pleading stage for failure to state a claim. To do so, the FJC searched the entire set using these different terms and phrases: "facts sufficient"; "sufficient facts"; "plausible claim"; "fails to state a claim"; "failed to state a claim"; "failing to state a claim" and "12(b)(6)."[96] Though a comprehensive search, it is possible that some motions to dismiss were missed that would have been found had broader search terms been tried.[97]

Even if the FJC missed filing activity equally (that is, both before *Twombly* and after *Iqbal*), underinclusiveness would still be highly consequential. Given the approach of the researchers in the study, which was to compare the total quantum of filings before *Twombly* to the total amount, after *Iqbal*, the size of the effect of the Court's cases turns on the amount of

---

[96] CECIL, ET AL. *supra* note 11, at 5 n.9.

[97] Unfortunately, the database the FJC used is not publicly available, so it is not possible to re-run the results. However, to illustrate how the FJC's search terminology may have led it to miss relevant motions, a search was run of a database in Westlaw that consists of federal pleadings and motions ("FED-FILING-ALL"). The following search was run: CO(TX) & "12(B)(6)" "FACTS SUFFICIENT" "SUFFICIENT FACTS" "PLAUSIBLE CLAIM" "FAILS TO STATE A CLAIM" "FAILED TO STATE A CLAIM" "FAILING TO STATE A CLAIM" & da(aft 9/2005 & bef 7/2006) % PRELIM("AMENDED MOTION" "SUPPLEMENT!" OPPOSITION RESPONSE REPLY! RECONSIDERATION OBJECTION (STRIKE /5 AFFIRMATIVE DEFENSES)). The search was limited to the same pre-*Twombly* time period the FJC studied but further limited only to federal district courts in Texas. Trying to run the search in all district courts produced more results than the 10,000 maximum of search results that Westlaw shows, so it was necessary to limit the search to less than all districts. No attempt was made to replicate the findings by looking at other districts.

Running the search using only the search terms and phrases the FJC used yielded 2,705 entries. The search was then run by adding these alternative terms: (FAIL! /5 STATE! /3 CLAIM! CAUSE! ACTION!). Broadening the search in this manner yielded substantially more entries (another 538, or 20% more than the previous yield). The same search was run in the post-*Iqbal* time period used by the FJC (October 2009-June 2010) and yielded 10% more than was produced by using the FJC terms only, not as great of a difference as in the pre-*Twombly* period but still a substantial number of additional motions. While not all were motions to dismiss for failure to state a claim, subsequent review found that most were. Results on file with author. Whether a similar broadening of the search terminology would have identified more relevant motions in the dataset that the FJC used is not known.

activity found. If there were twice as many motions actually filed as the FJC observed, then the size of *Twombly* and *Iqbal*'s effects would be twice as large as those that were reported, as noted above.[98]

CONCLUSION

Three primary assessments have been made of the FJC's study. Taking the last first, it was observed that there are reasons to be concerned that the study may be providing us an incomplete picture of actual Rule 12(b)(6) activity, especially as to the filing rate. Some of the possible explanations for underinclusiveness (such as the choice to look at only a 90-day window to find motions that were filed) may have biased the results. Even if the failure to capture all relevant motion activity was a non-biased error, the inclusiveness problem is consequential. Because the study was designed to compare over time the filing and grant rate of Rule 12(b)(6) motions, the size of the effect of the Court's cases turns on the amount of activity found.

Leaving to one side that the collected data may be incomplete, I argued that by emphasizing only whether the effects they observed were statistically significant, but not explaining what that technical terminology means, the study unintentionally confuses readers into thinking that the study proved *Twombly* and *Iqbal* were not responsible for the substantively significant changes in dismissal practices and outcomes that were found. Whether those changes were statistically significant or not, rulemakers could reasonably want to take into consideration that in the post-*Iqbal* period plaintiffs were twice as likely to face a dismissal motion, that judges have been granting more dismissal orders and, most importantly, that a defendant's chances of winning dismissal after *Iqbal* were better both overall and in every case category examined. The study could have aided transparency and understanding by clarifying what a no-significance finding means; discussing the cost of setting a high threshold for statistical significance; and raising plausible reasons with readers why the differences observed in dismissal outcomes might not have reached the defined level of statistical significance.

Even the foregoing is less than the entire story. Because of inherent limitations in doing empirical work of this nature, the cases may be having effects that the FJC researchers were unable to detect, as Part IV of the paper has shown. Comparing how many motions were filed and granted before *Twombly* to after *Iqbal* cannot tell us whether the Court's cases are deterring some claims from being brought, whether they have increased dismissals of complaints on factual-sufficiency grounds, or how many meritorious cases have been dismissed as a result of the Court's stricter pleading filter. Ultimately, then, perhaps the most important lesson to take away from this last assessment of the FJC's report is that empirical study cannot resolve all of the policy questions that *Twombly* and *Iqbal* raise.

---

[98] A third possibility is that some of the motions filed may not have been filed electronically or were otherwise not text-searchable. No further work has been done to determine how often this may occurred.