2012

Discovery about Discovery: Sampling Practice and the Resolution of Discovery Disputes in an Age of Ever-Increasing Information

Charles M. Yablon
Benjamin N. Cardozo School of Law, yablon@yu.edu

Nick Landsman-Roos

DISCOVERY ABOUT DISCOVERY: SAMPLING PRACTICE AND THE RESOLUTION OF DISCOVERY DISPUTES IN AN AGE OF EVER-INCREASING INFORMATION

Charles Yablon & Nick Landsman-Roos†

This Article provides the first extended academic consideration of a new practice adopted by an increasing number of courts to resolve e-discovery disputes—the sampling of a small portion of the information sought in backup or other relatively inaccessible files. We provide a comprehensive overview and statistical analysis of contemporary sampling techniques, identifying issues where sampling practice is inconsistent or where additional guidance appears to be required. Our aim is to provide a coherent theoretical approach to the use of sampling, suggesting "best practices" for many unresolved issues, and locating sampling practice within broader contemporary debates about discovery.

TABLE OF CONTENTS

INTRODUCTION 720

I. THE ROLE OF SAMPLING IN CONTEMPORARY DISCOVERY PRACTICE 724
- A. *The Rise and Decline of Full Disclosure* 724
- B. *The Origins and Current Law of Sampling* 732
- C. *A Normative Justification for Sampling* 736

II. STATISTICAL AND DESCRIPTIVE ANALYSIS OF REPORTED SAMPLING CASES 744
- A. *The Patchwork of Judicial Approaches* 745
- B. *How Courts Sample* 748
  - 1. Data Description and Methodology 748
  - 2. Results 753

III. NORMATIVE IMPLICATIONS OF THE ANALYSIS: RECOMMENDED "BEST PRACTICES" FOR SAMPLING 760
- A. *The Decision to Sample* 760

† The authors are, respectively, Professor of Law, Benjamin N. Cardozo School of Law and J.D. Candidate, Stanford Law School, 2013. We wish to thank Professor Richard Marcus for his very helpful suggestions on an early draft of this Article and Jil Simon for excellent research assistance.

B. *Method of Sampling*.....765
C. *Briefs Regarding Sampling*.....771
D. *Interpreting and Applying the Results of Sampling*.....772
CONCLUSION.....774
APPENDIX A: AN APPLICATION OF BAYES THEOREM TO SAMPLING DECISIONS.....774
APPENDIX B: PROPOSED "BEST PRACTICES" FOR THE USE OF SAMPLING IN THE RESOLUTION OF DISCOVERY DISPUTES.....776

INTRODUCTION

No subject in law is of more practical importance yet garners less theoretical attention than discovery in civil actions.[1] In most contemporary complex litigation, it is the primary focus of the pretrial process—the source of most of the cost, motion practice, and lawyer effort—that ultimately brings about a resolution of the dispute. That resolution, in turn, is usually a settlement based largely on the results and potential costs of the discovery process.[2] The advent of electronically stored data—so-called "e-discovery"—as a primary method of information storage and retrieval has resulted in discovery becoming even more complex, costly, and time consuming.[3]

While there is a large and growing literature containing guidance for practitioners and judges supervising the discovery process,[4] theoretical discussions of the problems raised by contemporary discovery practice remain extremely rare, both in academic journals and appellate court opinions.[5] Although the Supreme Court seems to have recently developed a strong interest in civil procedure—providing

---

1 Scott A. Moss, *Litigation Discovery Cannot Be Optimal but Could Be Better: The Economics of Improving Discovery Timing in a Digital Age*, 58 DUKE L.J. 889, 889–90 (2009).

2 *See generally* Robert D. Cooter & Daniel L. Rubinfeld, *An Economic Model of Legal Discovery*, 23 J. LEGAL STUD. 435 (1994) (discussing the effects of discovery on parties' incentives to settle).

3 In 2007, $2.70 billion was spent on e-discovery, representing a 43% increase in the amount spent in 2006. *See* George Socha & Tom Gelbmann, *A Look at the 2008 Socha-Gelbmann Survey*, LAW TECH. NEWS, Aug. 11, 2008. As of 2010 that figure has grown to $2.80 billion. *See* George Socha & Tom Gelbmann, *Climbing Back*, LAW TECH. NEWS, Aug. 1, 2010.

4 *See, e.g.*, Symposium, *Navigating the Changing Ethical and Practical Expectations for E-Discovery*, 36 N. KY. L. REV. 445, 445–629 (2009).

5 For example, in a recent symposium issue of the University of Denver Law Review, participants were asked to choose any Federal Rule of Civil Procedure and discuss proposals for its revision. Of the six law professors participating in the symposium, none choose to discuss a discovery rule. Of the three judge participants, one wrote about discovery. Of the five contributions by practitioners, three were about discovery rules. *See* Symposium, *Civil Justice Reform*, 87 DENV. U. L. REV. 213, 213–559 (2010). The theoretical literature that does relate to discovery has largely focused on cost allocation, as opposed to discussion of the ways in which courts should make such decisions. *See* Edward H. Cooper, *Discovery Cost Allocation: Comment on Cooter and Rubinfeld*, 23 J. LEGAL STUD. 465 (1994); Bruce L. Hay, *Effort, Information, Settlement, Trial*, 24 J. LEGAL STUD. 29 (1995); Cooter & Rubinfeld, *supra* note 2.

important new opinions in cases involving pleading standards, class actions, personal jurisdiction, and even *Erie* issues—the Justices have not seen fit to grant certiorari in any case primarily focused on civil discovery.[6] The academic literature, which provides extensive analyses and critiques of other aspects of the civil litigation process, also tends to give discovery short shrift.[7]

This is unfortunate, not just because of the practical importance of discovery, but also because concerns about the extent, cost, and burden of discovery lurk just below the surface in many recent Supreme Court decisions. For example, the controversial heightened pleading standard articulated in *Twombly*[8] and *Iqbal*[9] is expressly designed to protect certain defendants from the "burdens of discovery,"[10] which are said to be "sprawling, costly and hugely time-consuming."[11] Critics of those decisions decry their curtailment of "broad discovery," said to be one of the "integral, interdependent elements of the pretrial process."[12] The Supreme Court's recent decision in *Wal-Mart Stores, Inc. v. Dukes*,[13] which has been criticized for its extensive consideration of the merits on

6 A review of the discovery chapters of leading civil procedure casebooks reveals that the most recent Supreme Court decision relating to discovery is *Mohawk Industries, Inc. v. Carpenter*, 558 U.S. 100 (2009), which concerned the appealability of attorney-client privilege determinations, not the scope of discovery. *See* JACK H. FRIEDENTHAL, ARTHUR R. MILLER, JOHN E. SEXTON & HELEN HERSHKOFF, 2011–2012 CIVIL PROCEDURE SUPPLEMENT 571 (2011). The second most recent case is *Upjohn Co. v. United States*, 449 U.S. 383 (1981), another case about evidentiary privileges rather than the discovery rules. *See* JACK H. FRIEDENTHAL, ARTHUR R. MILLER, JOHN E. SEXTON & HELEN HERSHKOFF, CIVIL PROCEDURE: CASES AND MATERIALS 908 (10th ed. 2009); LINDA J. SILBERMAN, ALLAN R. STEIN & TOBIAS BARRINGTON WOLFF, CIVIL PROCEDURE: THEORY AND PRACTICE 625 (3d ed. 2009); *see also* Joel Slawotsky, *Rule 37 Discovery Sanctions—The Need for Supreme Court Ordered National Uniformity*, 104 DICK. L. REV. 471, 471 (2000) ("Nearly a quarter century has elapsed since the Supreme Court last addressed in a significant fashion discovery sanctions under Rule 37 of the Federal Rules of Civil Procedure.").

7 *See* Moss, *supra* note 1, at 893 ("Academics rarely focus on *how courts decide* discovery disputes (which, unlike trials, occur in most lawsuits), frustrating judges and parties alike." (emphasis added)). Professor Richard Marcus provides a possible explanation for this relative lack of academic attention: a widespread acceptance among academics of the "Liberal Ethos," the normative principle "that suits should be decided on their legal (substantive) merits and that procedure should be a Handmaid in that process." Richard Marcus, *Not Dead Yet*, 61 OKLA. L. REV. 299, 302 (2008). From this perspective, the changes made in the discovery rules since 1970 appear "retrograde" and, coupled with a general academic skepticism of practitioners' charges of pervasive discovery abuse, have led to a general tenor of disapproval regarding more recent attempts to limit discovery. *Id.* at 304–06. As the argument that "it wasn't broke and didn't need fixing" is one of the less challenging forms of academic discourse, it is perhaps not surprising that discovery has attracted less academic attention than other procedural issues.

8 Bell Atl. Corp. v. Twombly, 550 U.S. 544 (2007).

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

10 *Id.*

11 *Twombly*, 550 U.S. at 560 n.6.

12 Arthur R. Miller, *From* Conley *to* Twombly *to* Iqbal*: A Double Play on the Federal Rules of Civil Procedure*, 60 DUKE L.J. 1, 5 (2010).

13 131 S. Ct. 2541 (2011).

a class certification motion,[14] was justified, in part, by the very extensive discovery that had already taken place in connection with that motion.[15] And, in *Smith v. Bayer Corp.*,[16] although the Court ultimately rejected the attempt to enjoin a state court class action based on the issue preclusive effect of a prior federal denial of class certification, Justice Kagan acknowledged that the "strongest argument" for such preclusion is the cost of such serial relitigation and its attendant pretrial practices.[17]

So is discovery an expensive, burdensome thing whose abuses must be curbed, and coercive power minimized, whenever possible?[18] Or is it an essential right and integral part of modern litigation practice,[19] enabling parties and courts to develop the information they need to resolve an underlying dispute? The answer is surely both, and recent changes in discovery practice and the Federal Rules make it increasingly imperative that judges find fair, transparent, and effective means for resolving discovery disputes in individual cases. Unfortunately, because of the relative lack of academic or appellate court guidance, trial courts do not have a good theoretical framework with which to analyze discovery issues in different litigation situations or to balance the various factors that are material to adjudication of discovery disputes under the Federal Rules. Accordingly, magistrate and district court judges have been forced to develop rules and practices on a largely ad hoc basis.

This Article assists in developing a conceptual framework for contemporary discovery by providing the first theoretical and empirical analysis of a relatively new approach to resolving discovery disputes—the sampling of a small portion of the requested information prior to ruling on the underlying dispute. Sampling is expressly endorsed in the Advisory Committee notes to the 2006 revisions to Federal Rule of Civil Procedure 26(b)(2)(B), one of the so-called "e-discovery amendments."[20] Even before that time, it was applied in a number of influential discovery cases, most notably district court Judge Shira

---

14 *See generally* Alexandra D. Lahav, *The Case for "Trial by Formula,"* 90 TEX. L. REV. 571 (2012); Suzette M. Malveaux, *How Goliath Won: The Future Implications of* Dukes v. Wal-Mart, 106 NW. U. L. REV. COLLOQUY 34, 38 (2011).

15 There had been over one hundred and seventy-five depositions taken and more than a million pages of documents and electronic personnel data produced in the class certification stage in the court below. *See* Brief in Opposition of Respondents at 10–11, *Wal-Mart*, 131 S. Ct. 2541 (No. 10-277), 2010 WL 4220519 at *3.

16 131 S. Ct. 2368 (2011).

17 *Id.* at 2381.

18 For a recent critique of the current model of discovery cost allocation in which the producing party bears the expenses associated with its opponent's discovery requests, see Martin H. Redish & Colleen McNamara, *Back to the Future: Discovery Cost Allocation and Modern Procedural Theory*, 79 GEO. WASH. L. REV. 773 (2011).

19 *See* Geoffrey C. Hazard, Jr., *From Whom No Secrets Are Hid*, 76 TEX. L. REV. 1665, 1694 (1998) (arguing that discovery has achieved near-constitutional status in the United States).

20 FED. R. CIV. P. 26(b)(2)(B) is titled "Specific Limitations on Electronically Stored Information."

Scheindlin's seminal discussion of e-discovery practice in *Zubulake v. UBS Warburg LLC*.[21] It has become an increasingly popular technique for adjudication of discovery disputes, particularly in complex litigation with asymmetric discovery obligations[22] involving large amounts of electronically stored information. Since 1999, there have been at least forty reported cases in which sampling has been considered or utilized.[23]

This Article provides both an empirical and normative study of those cases, and of contemporary sampling practice generally. We find increasing acceptance and use of the practice, but also a wide variety of different approaches and conflicting methodologies. Courts have differed over when it is appropriate to use sampling techniques, the size of the sample to be used, the method by which the sample is to be selected, whether sampling should be used to cut-off further discovery or merely to shift costs, and many other issues. The Federal Rules provide no guidance on any of these matters and the discussion of these issues in the cases themselves is limited.

In order to identify issues where sampling practice is inconsistent or where additional guidance appears to be required, this Article provides the first comprehensive overview and statistical analysis of the practice. Our aim is to provide a coherent theoretical approach to the use of sampling, suggesting "best practices" for many unresolved issues and locating sampling practice within broader contemporary debates about discovery. We view sampling as a creative judicial response to the task created for the courts by the substitution of "proportionality" for "full disclosure" as the governing standard for discovery in contemporary civil litigation.

We believe sampling provides useful "discovery about discovery." By giving a court additional information about the data or documents being sought, sampling enables a court to make a better informed and more nuanced decision on the underlying discovery dispute. By requiring the court to rule coherently on such concepts as the "likely benefit"[24] of the proposed discovery or the "needs of the case,"[25] the proportionality standard requires the court to develop a much finer-grained understanding of the merits of the underlying litigation during the pretrial process. We believe properly conducted sampling can aid

---

21 229 F.R.D. 422 (S.D.N.Y. 2004) (*Zubulake V*). This case resulted in numerous written opinions, several of which will be referenced in this Article, according to their canonical desigantions. *See id.* at 424 ("This is the fifth written opinion in this case, a relatively routine employment discrimination dispute in which discovery has now lasted over two years."); *see also, e.g.*, Zubulake v. UBS Warburg LLC (*Zubulake I*), 217 F.R.D. 309 (S.D.N.Y. 2003); Zubulake v. UBS Warburg LLC (*Zubulake III*), 216 F.R.D. 280 (S.D.N.Y. 2003).

22 *See* discussion *infra* Part I.B; *see also* Rodney A. Satterwhite & Matthew J. Quatrara, *Asymmetrical Warfare: The Cost of Electronic Discovery in Employment Litigation*, 14 RICH. J.L. & TECH. 9 (2008).

23 *See infra* Part II.A.

24 FED. R. CIV. P. 26(b)(2)(C)(iii).

25 *Id.*

the court in its development of such an understanding, and should be applied in a fair, neutral, and cost-effective way, while recognizing that some prejudgment of the merits is a necessary consequence of the application of the proportionality standard.

This Article is divided into three parts. Part I is a conceptual overview that considers (a) the recent changes in discovery practice under the Federal Rules and the rise of proportionality as the governing standard, (b) the origins of sampling and the current legal rules governing it, and (c) a normative justification of sampling practice in enabling courts to apply the proportionality standard more fairly and effectively. Part II is an empirical study of the reported sampling cases, using both descriptive and statistical approaches to analyze the case law. It reviews in what contexts courts utilize sampling, the disparate approaches courts have taken in employing the sampling methodology, and the effect sampling has on the decision to shift costs or cut-off discovery. Part III seeks to apply the theoretical perspective of Part I to the practical questions revealed in Part II, with the goal of providing a new, relatively comprehensive account of "best practices" regarding when and how sampling techniques should be utilized.

We derive from this analysis some new and perhaps surprising conclusions. We discover from the existing case law a fairly strict upper limit on sampling of 25% of the total cost of the discovery sought. We believe that a presumptive sample size of between 15% and 25% is likely to provide maximum information to courts at minimum cost in most cases. We were also somewhat surprised to find that most courts do not follow a randomized method of "scientific sampling" but more frequently follow versions of a "best case scenario" approach in which the party seeking discovery gets to choose which files will be sampled. Although this approach has been implicitly criticized by some commentators who espouse scientific sampling, we believe it can, in fact, be justified in many cases as the optimal sampling methodology. Finally, although we find the courts have sharply split over whether sampling should be used to cut-off discovery or merely shift costs, we provide a normative rule under which both results can be justified in different circumstances based on the information provided by sampling.

I. THE ROLE OF SAMPLING IN CONTEMPORARY DISCOVERY PRACTICE

A. *The Rise and Decline of Full Disclosure*

Discovery is not what it used to be. The drafters of the Federal Rules, who created modern federal discovery practice,[26] were motivated

[26] For discussion of the "revolutionary" change represented by the discovery provisions of

by a vision of "full disclosure" aimed at eliminating surprises at trial[27] and ensuring that adjudication would be based on the fullest possible evidentiary record.[28] Mutual self-interest would cause the parties to exchange relevant information about the critical factual allegations of the case in what would be a largely self-regulating process.[29] This would lead to fairer, more efficient, and shorter trials, as well as more and better settlements.[30]

This vision has long since ceased to be the dominant view of how federal civil discovery should be conducted.[31] Having reached an apex of sorts with the 1970 revision to the Federal Rules, the ideal of full disclosure has been in retreat ever since. Concerns about over-discovery, "fishing expeditions," imposition of massive costs, and other alleged abuses of the discovery process by those claiming a right to "full disclosure" emerged shortly after 1970[32] and have led to repeated modifications and limitations of the discovery rules.[33] The trend of these changes has been to grant judges greater power to limit parties' access to information in the possession of their opponents. The standard of "proportionality" has been superimposed over the ideal of full

---

the original Federal Rules, see Geoffrey C. Hazard, Jr., *Discovery Vices and Trans-Substantive Virtues in the Federal Rules of Civil Procedure*, 137 U. PA. L. REV. 2237, 2238–39 (1989); Stephen N. Subrin, *Fishing Expeditions Allowed: The Historical Background of the 1938 Federal Discovery Rules*, 39 B.C. L. REV. 691, 734 (1998).

27 *See* Edson R. Sunderland, *Foreword* to GEORGE RAGLAND, JR., DISCOVERY BEFORE TRIAL, at iii (1932):

> False and fictitious causes and defenses thrive under a system of concealment and secrecy in the preliminary stages of litigation followed by surprise and confusion at the trial. Under such a system the merits of controversies are imperfectly understood by the parties, are inadequately presented to the courts, and too often fail to exert a controlling influence upon the final judgment.

Id.

28 After the 1970 revisions, a simple request was sufficient. Prior to that time, the party seeking relevant information had to show "good cause," a standard to which the current rules have partially returned. John H. Beisner, *Discovering a Better Way: The Need for Effective Civil Litigation Reform*, 60 DUKE L.J. 547, 560–61 (2010).

29 *Id.* at 557.

30 STEPHEN N. SUBRIN & MARGARET Y. K. WOO, LITIGATING IN AMERICA: CIVIL PROCEDURE IN CONTEXT 144 (2006); *see also* Subrin, *supra* note 26, at 716.

31 Yet how well the current system is functioning remains a matter of considerable dispute. *See* EMERY G. LEE III & THOMAS E. WILLGING, FED. JUDICIAL CTR., NATIONAL, CASE-BASED CIVIL RULES SURVEY 27 (2009), *available at* http://www.fjc.gov/public/pdf.nsf/lookup/dissurv1.pdf/$file/dissurv1.pdf ("Respondents were asked to rate the information generated by the parties in discovery in the closed case. . . . Both plaintiff and defendant attorneys tended to answer 'just the right amount;' 56.6 and 66.8 percent, respectively, gave that answer.").

32 *See, e.g.*, Richard Marcus, *Retooling American Discovery for the Twenty-First Century: Toward a New World Order?*, 7 TUL. J. INT'L & COMP. L. 153, 155 (1999); Edward F. Sherman & Stephen O. Kinnard, *Federal Court Discovery in the 80's—Making the Rules Work*, 95 F.R.D. 245, 246 (1982); Jeffrey W. Stempel & David F. Herr, *Applying Amended Rule 26(B)(1) in Litigation: The New Scope of Discovery*, 199 F.R.D. 396, 401–02 (2001).

33 Since their promulgation, the discovery rules have been amended twelve times: in 1946, 1963, 1966, 1970, 1980, 1983, 1987, 1993, 2000, 2006, 2007, and 2010. FED. R. CIV. P. 26.

disclosure.[34] The right to relevant information must be weighed against the stakes in the litigation, the costs of obtaining it, and the benefit it is likely to provide to the party seeking it. These concerns are exacerbated when the information sought is not in "reasonably accessible" electronic form.[35]

These changes to discovery rules reflected a series of broader technological, social, and legal changes. The first has been an immense increase in the amount of potentially discoverable information. In 1970, duplicating machines were just beginning to be common office equipment, making every copy with some handwritten notes in the margin a "non-identical copy" which had to be separately searched for and disclosed.[36] Fax machines and electronic data storage also expanded the universe of discoverable data. The biggest change, however, has been the advent of e-mail. Conversations that once would have occurred over the telephone or at the water cooler—and would have therefore vanished without a trace—are now being preserved and retained, perhaps indefinitely, in the data storage of the company.

An equally important legal development has been an enormous increase in "asymmetric litigation"—cases in which one side possesses almost all the relevant information. As Judge Richard Posner recently observed, in a class action he dubbed "nearly frivolous":

> In most class action suits, including this one, there is far more evidence that plaintiffs may be able to discover in defendants' records (including emails, the vast and ever-expanding volume of which has made the cost of discovery soar) than vice versa. For usually the defendant' conduct is the focus of the litigation and it is in their records, generally much more extensive than the plaintiffs' (especially when as in a consumer class action the plaintiffs are individuals rather than corporations or other institutions), that the plaintiffs will want to rummage in quest of smoking guns.[37]

This change in the conception of "typical" litigation from one in which each side has relevant information to one in which discovery is a weapon to be wielded by one party against the other underlies much of the effort to roll back the scope of discovery under the Federal Rules.[38]

---

34 This standard of "proportionality" was codified in FED. R. CIV. P. 26(b)(2)(C)(iii).

35 *See* FED. R. CIV. P. 26(b)(2)(B).

36 *See* Metro. Opera Ass'n v. Local 100, Hotel Emps. & Rest. Emps. Int'l Union, 212 F.R.D. 178, 181 (S.D.N.Y. 2003) (explaining that drafts, non-identical copies, and electronic copies all must be disclosed).

37 Thorogood v. Sears, Roebuck & Co., 624 F.3d 842, 849–50 (7th Cir. 2010), *cert. granted and judgment vacated*, 131 S. Ct. 3060 (2011) (mem.). The growth of personal computers with their own difficult-to-completely-erase cache of discoverable e-mails belonging to plaintiffs may someday restore some reciprocity and "two-sidedness" to the discovery process, but that day still seems far off.

38 It is worth remembering that a private right to sue for federal securities law violations was only clearly established in 1964, *see* J.I. Case Co. v. Borak, 377 U.S. 426, 432 (1964), which happens to be the same year Title VII of the Civil Rights Act of 1964, Pub. L. No. 88-352, 78

Closely connected to the development of these new kinds of claims (and the rise of a specialized group of plaintiffs' counsel to prosecute them) is a third, subtler, and perhaps more controversial change in the concept of what constitutes relevant information. Whereas traditional litigation turned largely on development and exposition of concrete facts about the actions of the parties and the events giving rise to the claims, a critical question in much of the new asymmetric litigation against corporations turns on the knowledge or beliefs of the companies' agents. So the question of whether a company knew a statement in its securities filing had become untrue, or had notice of the dangers or addictive nature of their product, or fired a worker with discriminatory intent, are all matters in which the difference between winning and losing the case would appear to be the contents of e-mails and other documents found in the companies files.[39] In such cases, access to broad ranging discovery will seem critical to plaintiffs yet may appear to defense counsel as an unguided "fishing expedition."

These changes have made it substantially more difficult to realize the original drafters' vision of discovery as a self-regulating cooperative process.[40] When certain parties, and the lawyers who represent them, know that they will consistently be either requesting or resisting discovery, they have an incentive to take maximalist positions with respect to such disputes. To be sure, professional ethics and the discovery rules place some limits on the extremity of positions lawyers

---

Stat. 253 (codified as amended at 42 U.S.C. §§ 2000e–2000e-14 (2006)), the basis for most employment discrimination class actions, was passed. Most consumer class actions came even later, and many products liability claims, although still not able to be brought as class actions, are large complex lawsuits where discovery is strongly asymmetric. *See* John Cirace, *A Theory of Negligence and Products Liability*, 66 ST. JOHN'S L. REV. 1, 63–64 (1992); Satterwhite, *supra* note 22, at 6–9 (discussing asymmetrical litigation in the employment discrimination context); Wendy E. Wagner, *What's It All About, Cardozo?*, 80 TEX. L. REV. 1577, 1592–93 (2002).

39 Consider the following statement from the website of an electronic data discovery support firm, which adopts such a perspective in an ad aimed at plaintiffs' counsel:

> Every litigation team is looking for it. The "smoking gun." You know where it is—it's in that forest of paper that opposing counsel has—and you've got to find it. But what if the "gun" isn't paper, and never was? What if there is no forest? It may just be made up of bytes, zeroes and ones that mean something electronically. Today, more than 93 percent of documents generated are originally created in an electronic format. . . . As such, a traditional discovery demand will not capture this electronic information if you don't specifically ask for it. . . . The scariest prospect is the sheer amount of data created, and it is nothing short of infinite. In 2008, internet users alone, will generate just over one hundred sixty-one (161) exabytes of data (an exabyte is a billion gigabytes). The average American can have more than a gigabyte of just e-mail on his or her hard drive, any one of which may be highly relevant to your case.

Robert Almoney, *Smoking Guns: Electronic Intentions, Data and Discovery Costs*, LEGALIS, http://www.legalis.com/WhitePapers-SmokingGuns (last visited Oct. 20, 2011).

40 John S. Beckerman, *Confronting Civil Discovery's Fatal Flaws*, 84 MINN. L. REV. 505, 515–16 (2000); Steven S. Gensler, *Some Thoughts on the Lawyer's E-volving Duties in Discovery*, 36 N. KY. L. REV. 521, 530, 540 (2009).

can responsibly take, and lawyers may feel a responsibility to honor the various calls for cooperation in discovery made implicitly in the discovery rules and explicitly by critics of the existing system.[41] The fact remains, however, that lawyers engaged primarily in asymmetric litigation tend to view discovery disputes from a narrower perspective than those who are familiar with both sides of such disputes.[42] This tendency toward polarization has extended beyond litigation to law reform efforts as well.[43]

In light of powerful calls for either radical change or no change at all, the gradual development of a complex and nuanced rule that seeks to accommodate both of these perspectives must be judged a single achievement of the rulemaking process. The product of that process, the "proportionality" standard, has now become the prevailing legal criterion for resolving discovery disputes.[44] While the basic scope of

---

41 *See* The Sedona Conference, *The Sedona Conference Cooperation Proclamation*, 10 SEDONA CONF. J. 331, 331 (2009).

42 *See, e.g.*, Thomas E. Willging et al., *An Empirical Study of Discovery and Disclosure Practice Under the 1993 Federal Rule Amendments*, 39 B.C. L. REV. 525, 540 (1998) (finding that plaintiff attorneys were more likely to complain that a party failed to respond adequately, while defense attorneys lamented the vagueness of requests or number of documents sought). *But see* James S. Kakalik et al., *Discovery Management: Further Analysis of the Civil Justice Reform Act Evaluation Data*, 39 B.C. L. REV. 613, 638 (1998) (finding no statistical difference in the percentage of lawyers who viewed the management of discovery as fair).

43 Groups that see themselves as primarily representing corporate interests have been a strong voice for greater restrictions on discovery, while those who represent plaintiffs' interests point to the negative impact of such changes. *See* JOHN H. BEISNER, U.S. CHAMBER OF COMMERCE INST. FOR LEGAL REFORM, THE CENTRE CANNOT HOLD: THE NEED FOR EFFECTIVE REFORM OF THE U.S. CIVIL DISCOVERY PROCESS, 18–19 (2010) ("[C]orporations . . . now list discovery as their most pressing concern when litigation is imminent."); Joshua M. Koppel, *Tailoring Discovery: Using Nontranssubstantive Rules to Reduce Waste and Abuse*, 161 U. PA. L. REV. (forthcoming 2012), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2154468 (manuscript at 20) ("Some scholars and practitioners, mostly defendants' organizations, argue that parties are able to use the broad tools of discovery to impose costs on their adversaries that push the adversary toward settlement."). On the other hand, plaintiffs' attorneys feel the need to use "the discovery rules in a way that either brings relevant information to light, or assures that a party that suppresses information bears the consequences." Stuart Ollanik, *Wielding the Tools of Discovery*, AMERICAN ASS'N FOR JUSTICE, Fall 2007, at 8 (listing various methods to best use the discovery process to the plaintiff's benefit).

44 *See* 8 CHARLES ALAN WRIGHT, ARTHUR R. MILLER, MARY KAY KANE & RICHARD L. MARCUS, FEDERAL PRACTICE AND PROCEDURE § 2008.1 (3d ed. 2010). "[W]hen the second edition of [Federal Practice and Procedure] appeared in 1994, it seemed that the Rule 26 proportionality provision had . . . produced only a ripple in the caselaw" and some courts continued to address discovery issues "without invoking the new language." *Id.* Since then, however, "attention to the proportionality provisions has grown . . . and endorsement of their use has widened." *Id.*; *see also* Ronald J. Hedges, *A View from the Bench and the Trenches: A Critical Appraisal of Some Proposed Amendments to the Federal Rules of Civil Procedure*, 227 F.R.D. 123, 126–27 (2005); Richard L. Marcus, *Discovery Containment Redux*, 39 B.C. L. REV. 747, 773 (1998); Richard L. Marcus, *Retooling American Discovery for the Twenty-First Century: Toward a New World Order?*, 7 TUL. J. INT'L & COMP. L. 153, 162–63 (1999); Shira A. Scheindlin & Jeffrey Rabkin, *Electronic Discovery in Federal Civil Litigation: Is Rule 34 Up to the Task?*, 41 B.C. L. REV. 327, 349 (2000); Panel Discussion, *Managing Electronic Discovery: Views from the Judges*, 76 FORDHAM L. REV. 1, 24 (2007). However, "attention to the proportionality

disclosure under Federal Rule of Civil Procedure 26(b)(1) remains almost as broad as it was in 1970, Rule 26(b)(2)(C) instructs federal judges that they "must limit" the frequency or extent of otherwise permitted discovery if they determine that (i) it is unreasonably cumulative or may be obtained from a less burdensome, more convenient source; (ii) the party has already had "ample opportunity" to obtain the information sought; and (iii) it prescribes an explicit balancing test in which the "burden and expense of the proposed discovery" is weighed against its "likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues."[45]

Those same considerations are referenced in Rule 26(b)(2)(B), a more recent limitation on discovery enacted as part of the December 1, 2006 e-discovery amendments to the Federal Rules.[46] Rule 26(b)(2)(B) requires the responding party to show that the electronically stored information requested is "not reasonably accessible because of undue burden or cost."[47] Such information may still be obtained, but only upon a showing of "good cause" by the requesting party, and good cause is expressly made subject to the proportionality considerations of Rule 26(b)(2)(C).[48]

The strength of the proportionality standard is that it is able to maintain the rules drafters' fundamental vision of wide-ranging discovery, the costs of which are borne primarily by the producing party, while recognizing that such a vision is subject to abuse and must be carefully supervised to avoid becoming a device for unnecessarily burdening or coercing the party against whom discovery is sought. The proportionality standard guards against such abuses, however, not by restricting the scope or availability of discovery at the outset, but by subjecting potentially burdensome requests to a searching, individualized analysis of their costs and benefits in the particular litigation within which they are made. The problem is that this approach requires that complex and searching analysis from a neutral decision-maker, i.e., a judge or magistrate judge, who must develop a clear and

provisions has grown since 1994, and endorsement of their use has widened." WRIGHT ET AL., *supra* § 2008.1.

45 FED. R. CIV. P. 26(b)(2)(C)(i)–(iii).

46 Supreme Court of the United States, Order Adopting Amendments to the Federal Rules of Civil Procedure ¶ 3 (Apr. 12, 2006), *available at* http:// www.supremecourt.gov/orders/courtorders/frcv06p.pdf.

47 FED. R. CIV. P. 26(b)(2)(B).

48 "Good cause" is to be determined by "considering the limitations of Rule 26(b)(2)(C)." FED. R. CIV. P. 26(b)(2)(B). Some commentators have argued that an exact "good cause" test remains elusive. *See* Rachel Hytken, *Electronic Discovery: To What Extent Do the 2006 Amendments Satisfy Their Purposes?*, 12 LEWIS & CLARK L. REV. 875, 890–91 (2008); Henry S. Noyes, *Good Cause Is Bad Medicine for the New E-Discovery Rules*, 21 HARV. J.L. & TECH. 49, 87–91 (2007).

coherent perspective on the case whose pretrial process is being supervised.

Consider how the proportionality standard plays out in the kind of asymmetric discovery cases previously described. Although the balancing test of Rule 26(b)(2)(B)(iii)[49] is stated as a neutral principle, it is clear that in the vast majority of cases, absent cost-shifting, the "burden and expense" of discovery is going to fall on the party against whom discovery is sought, the corporate defendant, while most, if not all of the "likely benefit" of the discovery is likely to inure to plaintiffs by helping them prove their case. Of the various "considerations" mentioned by the Rule, only two—the amount in controversy and the parties' resources—are likely to be relatively objectively ascertainable by the court without regard to merits-based considerations. Those considerations are likely to provide clear answers to only a small portion of discovery disputes. When the amount in controversy and defendants' resources are both substantial, however, and the proposed discovery, while expensive, is still only a small fraction of those amounts, then Rule 26(b)(2)(B)(iii) directs consideration of "the needs of the case," "the importance of the issues at stake in the litigation," and the "importance of the proposed discovery in resolving the issue."[50] It is likely that the plaintiffs seeking discovery and the defendants resisting it will have very different views on these matters, differences rooted in their distinct views of the underlying merits of plaintiffs' claims. Judges may be understandably reluctant or unable to resolve these merits-based disputes before ruling on the pending discovery issues.

The Advisory Committee notes to the 2006 amendments recognize this dilemma and explicitly authorize sampling as a way to deal with it. In describing the "good cause" determination of Rule 26(b)(2)(B), which expressly references the "limitations of Rule 26(b)(2)(C)," the Committee states:

---

49 The Advisory Committee notes produced a five-factor balancing test in connection with its promulgation of Rule 26(b)(2)(B), and various other tests can be derived from the cases predating the 2006 amendments. For a comparative analysis of such tests, see Robert E. Altman & Benjamin Lewis, Note, *Cost-shifting in ESI Discovery Disputes: A Five Factor Test to Promote Consistency and Set Party Expectations*, 36 N. KY. L. REV. 569, 587–92 (2009). Such tests include the "marginal utility" test articulated in *McPeek v. Ashcroft*, 202 F.R.D. 31, 34 (D.D.C. 2001), the *Rowe* eight-factor test set forth in *Rowe Entertainment, Inc. v. William Morris Agency, Inc.*, 205 F.R.D. 421, 428–31 (S.D.N.Y. 2002), the *Zubulake* factors, *see Zubulake I*, 217 F.R.D. 309, 321–22 (S.D.N.Y. 2003), and a modification to the *Zubulake* test in *Wiginton v. CB Richard Ellis, Inc.*, 229 F.R.D. 568, 572–73 (N.D. Ill. 2004). The ABA has also advocated its own sixteen-factor balancing test. *See* SECTION OF LITIG., AM. BAR ASS'N, CIVIL DISCOVERY STANDARDS, Standard 29(b)(iv)(A-P) (2004). We are skeptical as to whether these various tests make much difference to judges actually engaged in the application of the proportionality standard and agree with Judge Scheindlin, who has stated, albeit in different terms, that the most important considerations under any of these standards is the simple "marginal utility" test: whether the likely value of the requested information outweighs the likely burden and cost of producing it. *See Zubulake I*, 217 F.R.D. at 323.

50 FED. R. CIV. P. 26(b)(2)(B)(iii).

> The good-cause determination, however, may be complicated because the court and parties may know little about what information the sources identified as not reasonably accessible might contain, whether it is relevant, or how valuable it may be to the litigation. In such cases, the parties may need some focused discovery, which may include sampling of the sources, to learn more about what burdens and costs are involved in accessing the information, what the information consists of, and how valuable it is for the litigation in light of information that can be obtained by exhausting other opportunities for discovery.[51]

The Advisory Committee makes a significant distinction about the information that may be obtained by sampling. It may tell the court not just whether the requested discovery is "relevant," but more importantly, "how valuable it is for the litigation." Implicit in this distinction is recognition that not all relevant information is equally valuable, or equally subject to discovery. Indeed, under modern concepts of relevance and the broader readings of the Rule 26(b)(2) standard, even the absence of information in defendants' files may have some "relevance" to the disputed claims. The proportionality test, however, presumes that relevance is not a binary property, but one that, like the cost and burden against which it is measured, can come in different degrees. When Rule 26(b)(2)(C)(iii) requires an evaluation of the "likely benefit" of the discovery and its "importance in resolving the issues," it is asking a court to determine not merely the relevance of the information sought, but its value to the litigation as a whole. This requires a fuller, more probing inquiry into the nature of the information being sought, information that frequently can only be provided in a cost-effective way by sampling.

There remains, however, the question of weight. Even when it is shown that the requested discovery is likely to contain not just relevant but "valuable" or "important" information, can it be outweighed by the countervailing considerations of burden and expense? The appropriate weighting of these various considerations is not specified under the Rule. Are a plaintiff's need and a defendant's burden equal considerations, or does a substantial showing of need by a plaintiff prevail against an equally substantial showing of burden by a defendant? Here we think it is important to recognize that the principle of proportionality is not a replacement for the principle of full disclosure, but an amendment and limitation on that principle, which continues to be the basic philosophy of the federal discovery rules. The point is well made by Professor Richard Marcus:

> Overall, [post-1970 rule changes providing for proportionality review and other limitations on federal discovery] represent[] a significant

---

51 FED. R. CIV. P. 26 advisory committee's note, 2006 amend.

> retrenchment from the broadest views of discovery that emerged in the 1960s. At the same time, it is important to appreciate that there was no renunciation of the basic idea of broad discovery, or the general notion that the responding party cannot force the other side to pay the cost of discovery.[52]

The clear implication of this view is that when files are likely to contain information that is valuable and important in proving the plaintiff's case, discovery of such information will be ordered, and at defendant's expense.[53] But how is a judge supposed to make a determination when the contents of defendants' files remain unknown, not just to the court and the plaintiffs, but very possibly to the defendants as well? It is in such circumstances that sampling may provide critical information to the court.

B. *The Origins and Current Law of Sampling*

Although sampling has mostly occurred in the federal courts,[54] the first reported discussion of sampling came from a Massachusetts state court, *Linnen v. A.H. Robins Co.*[55] It was one of numerous product liability suits filed following the detection of a link between the commonly prescribed appetite suppressant "fen-phen" and valvular

---

52 Richard Marcus, Essay, *Only Yesterday: Reflections on Rulemaking Responses to E-Discovery*, 73 FORDHAM L. REV. 1, 7 (2004). Professor Marcus is Special Reporter to the Discovery Subcommittee of the Judicial Conference Advisory Committee on Civil Rules, 1996–present. In this Essay, Professor Marcus also noted that the e-discovery amendments were designed to preserve the same balance as prior rule changes between the wide availability of broad discovery and the power of the courts to curb abuses. *Id.* at 13–14.

53 This is also consistent with Judge Scheindlin's weighting of the relevant cost-shifting factors in *Zubulake III*, though she warns the "list of factors is not merely a matter of counting and adding; it is only a guide." 216 F.R.D. 280, 289 (S.D.N.Y. 2003). The first two factors, "(1) the extent to which the request is specifically tailored to discover relevant information;" and "(2) the availability of such information from other sources . . . comprise the 'marginal utility test'" and "should be weighted the most heavily in the cost-shifting analysis." *Id.* at 284. Factors three, four, and five "address[] cost issues." *Id.* at 287. Noting that "[i]n an ordinary case, a responding party should not be required to pay for the restoration of inaccessible data if the cost of that restoration is significantly disproportionate to the value of the case," Judge Scheindlin found these factors to weigh against cost-shifting in *Zubulake*, because of the case's potential value. *Id.* at 288. Judge Scheindlin stated that factor six will often be "neutral" because litigation rarely presents a "novel issue." *Id.* at 288–89. Finally, factor seven, was found to "weigh[] in favor of cost-shifting" because the defendant "would not restore any of th[e] data of its own volition." *Id.* at 289.

54 In the course of our research, we only came across seven cases in which a state court used or considered the use of a sampling technique in the electronic discovery context. Of course, since most state trial court decisions, as well as those of many federal district courts, are never published (and that is likely to be particularly true with respect to decisions on discovery disputes), there may be considerably more court decisions concerning sampling that are simply unavailable for review.

55 No. 97-2307, 1999 WL 462015, at *5–7 (Mass. Super. Ct. June 16, 1999).

heart disease.[56] In *Linnen*, the plaintiffs sought production of e-mail messages retained by one of the defendants, Wyeth-Ayerst Laboratories (Wyeth). Wyeth opposed this request as duplicative,[57] costly, and burdensome.[58]

The plaintiffs' management committee reached an agreement with Wyeth whereby Wyeth would restore a sample of the backup tapes.[59] Wyeth had agreed to bear the initial costs of restoring the sample, but had the right to seek reimbursement.[60] This agreement was reported to the Massachusetts Superior Court, which approved it, stating that the "tapes have the potential for containing relevant material and that plaintiffs should have the opportunity to examine at least a portion of the tapes to determine if that is the case."[61] The *Linnen* case contains no discussion of the techniques to be used in determining the sample or how the parties came up with the sampling technique in the first place.

The first federal opinion discussing the use of sampling was *McPeek v. Ashcroft*.[62] Steven McPeek, an employee in the Bureau of Prisons, claimed he was sexually harassed by his then-supervisor and was subsequently retaliated against for filing harassment claims.[63] After the Department of Justice produced paper and electronic documents, McPeek requested a search of the Department's backup system since it might yield data deleted by computer users.[64] The Department protested that the possibility of relevant evidence from such a search was remote and could not justify the cost.[65]

---

56 *See In re* Diet Drugs, 582 F.3d 524, 529 (3d Cir. 2009).

57 Some of the earliest "fen-phen" litigation took place in Texas state courts. Those lawsuits generated expansive discovery; for example, Wyeth produced approximately three million pages of documents. *Id.* at 529–30.

58 *Linnen*, 1999 WL 462015, at *1. Wyeth maintained a software system that backed up intra-office communications on a daily basis. Those backup tapes also contained word processing files, spreadsheets, models, and e-mails saved on Wyeth's computers. The tapes were intended to be used only for data recovery in the event of a catastrophic disaster, and therefore were only kept for three months and then recycled. In 1997, Wyeth suspended its usual recycling practices and saved all backup tapes. A year later, Wyeth discovered a number of backup tapes from January 1994 to May 1995 that had been saved during the pendency of unrelated litigation. It was those tapes that were the subject of the plaintiffs' motion to compel. *Id.* The court estimated that 746 backup tapes could potentially contain relevant information, and estimates of the cost of restoring those tapes ranged from $1.15 to $1.75 million. *Id.* at *4. The plaintiffs in the multi-district litigation sought the same backup tapes. *Id.*

59 *Id.* at *5.

60 *Id.* Further discovery beyond that sample would be allowed only upon a showing of "good cause."

61 *Id.* at *6. The court reserved "any decision to require additional tapes to be restored until the potential for relevant and responsive documents has been more fully explored through review of the restored sample tapes." *Id.*

62 202 F.R.D. 31 (D.D.C. 2001).

63 *Id.* at 31–32.

64 *Id.* at 32.

65 *Id.* The Department's backup system was designed to permit recovery from a disaster, not archival preservation, and accordingly there were backup tapes for some periods but not others. *Id.*

Faced with the conflict between the "theoretical possibility" that a smoking gun might be hidden in the government's backup tapes, and the very real cost of restoring those tapes in monetary terms and human hours, the court elected to "take small steps and perform, as it were, a test run."[66] Magistrate Judge Francis Facciola set the parameters of this "test run": the defendant would perform a backup tape restoration of the e-mails attributable to a period selected by the court, based on its understanding of when the allegedly wrongful conduct was most likely to have occurred.[67] The defendant was to document the time and money spent doing the search, and upon completing the search the court would balance the search results against the expenses to determine whether additional discovery should go forward.[68]

Two years later, Judge Scheindlin fleshed out the sampling approach in her seminal *Zubulake* opinions.[69] *Zubulake* was an action by a former employee against her employer for retaliation and gender discrimination.[70] In response to Zubulake's discovery requests, UBS produced approximately 350 pages of documents, including approximately 100 pages of e-mails.[71] UBS, however, never searched for responsive e-mails on its backup tapes because it told plaintiff the cost of producing e-mails would be prohibitive (estimated at approximately $300,000).[72] UBS further argued that the cost of any discovery of the backup tapes should be shifted to Zubulake given the professed low probability that such a search would produce evidence worth the cost involved.[73]

After setting out a multi-factor test for cost-shifting, Judge Scheindlin expressed concern that courts applying cost-shifting analyses frequently made assumptions about the likelihood of finding relevant information without any factual basis.[74] This fact-less approach to cost-shifting was problematic for Judge Scheindlin because it is arguably inconsistent with Rule 26(b)(1)'s liberal approach to discovery, and

---

66 *Id.* at 33–34.

67 *Id.* at 34–35.

68 *Id.* at 35.

69 *Zubulake I*, 217 F.R.D. at 322–23.

70 *Id.* at 312.

71 *Id.* at 312–13.

72 *Id.* at 313. UBS later informed the court that the cost of restoring those e-mails on the backup tapes alone would cost approximately $175,000.00, exclusive of attorney time for reviewing the e-mails. Zubulake sought damages of approximately $13 million. *Id.* at 311 n.9, 311–12.

73 *Id.* at 313–14, 16.

74 Judge Scheindlin provided two examples of this approach: In *Rowe Entertainment, Inc. v. William Morris Agency, Inc.*, 205 F.R.D. 421, 430 (S.D.N.Y. 2002), the court relied on the *lack* of witness testimony or documentary evidence "showing that the e-mails are likely to be a gold mine" to find *in favor* of cost-shifting. Similarly in *Murphy Oil USA, Inc. v. Fluor Daniel, Inc.*, No. 99-3564, 2002 WL 246439, at *5 (E.D. La. Feb. 19, 2002), the court ordered cost-shifting because "the marginal value of searching the e-mail is modest at best" and the plaintiff "has not pointed to any evidence that shows that 'the e-mails are likely to be a gold mine.'"

because as a practical matter it is rare that a plaintiff will have sufficient evidence *before* conducting discovery to demonstrate the necessity of specific discovery. Judge Scheindlin ordered limited sampling of the backup tapes in question in order to inform the cost-shifting analysis with factual evidence gleaned through the sampling. In her words:

> When based on an actual sample, the marginal utility test will not be an exercise in speculation—there will be tangible evidence of what the backup tapes may have to offer. There will also be tangible evidence of the time and cost required to restore the backup tapes, which in turn will inform the second group of cost-shifting factors. Thus, by requiring a sample restoration of backup tapes, the entire cost-shifting analysis can be grounded in fact rather than guesswork.[75]

Judge Scheindlin's opinion in *Zubulake* has since had an enormous influence not only on the use of sampling,[76] but also on cost-shifting in electronic discovery generally.[77]

We can see from these decisions that sampling developed as an attempt to apply proportionality analysis to the costs of restoring and searching backup tapes in situations involving asymmetric litigation. In each case the courts were faced with balancing the high costs of the requested production against substantial uncertainty regarding the importance of the requested information. The courts in each case also sought to reduce that uncertainty by informing themselves, through sampling, about the likely contents of the requested information as well as the actual costs of discovery.[78] Further, in each of the cases the court was not just concerned with whether the potential information was relevant, but whether it was *so* important that it constituted a "smoking gun"[79] or "gold mine"[80] sufficient to outweigh the cost of production.

In light of the foregoing developments, and the recognition of the sampling approach in the Advisory Committee notes to the 2006 e-

---

75 *Zubulake I*, 217 F.R.D. at 324.

76 A number of decisions have relied on *Zubulake* as precedent for the use of sampling. *See, e.g.*, Parkdale Am., LLC v. Travelers Cas. & Sur. Co. of Am., Inc., Civil No. 3:06CV78-R, 2007 WL 4165247, at *12–13 (W.D.N.C. Nov. 19, 2007); Quinby v. WestLB AG, 245 F.R.D. 94, 102 (S.D.N.Y. 2006); Hagemeyer N. Am., Inc. v. Gateway Data Scis. Corp., 222 F.R.D. 594, 602 (E.D. Wis. 2004).

77 *See* James M. Evangelista, *Polishing the "Gold Standard" on the E-Discovery Cost-Shifting Analysis:* Zubulake v. UBS Warburg, LLC, 9 J. TECH. L. & POL'Y 1, 3 (2004) ("[T]he Zubulake decisions will become the seminal reference on this issue."); Christopher L. Troy & Margaret K. Simpson, *"Electrifying" Changes to the Federal Discovery Rules*, 36 SPG BRIEF 32, 34 (2007) ("The multiple opinions of *Zubulake v. UBS Warburg* have received the most attention and have become touchstones for resolving electronic discovery disputes.").

78 *Zubulake I*, 217 F.R.D. at 312; McPeek v. Ashcroft, 202 F.R.D. 31, 34 (D.D.C. 2001); *see also supra* note 72.

79 *Zubulake I*, 217 F.R.D. at 311 n.8.

80 *Id.* at 323.

discovery amendments,[81] the "concept of sampling to test both the cost and the yield is now part of the mainstream approach to electronic discovery."[82] While sampling has made its way into the judicial toolbox for handling electronic discovery disputes, the ways in which sampling has been implemented have been inconsistent across courts, and there has been no attempt to explain how sampling fits into the ongoing debates about managerial judging and discovery abuse. That is the task of the next Section.

C. *A Normative Justification for Sampling*

Sampling is not a completely unprecedented departure from prior discovery practice. Rather, it is a logical extension of the trend toward "managerial judging"—the reconceptualization of the judicial role from that of a passive, disinterested umpire to that of an active participant exercising various forms of discretionary adjudicative power over the pretrial process. The move to managerial judging has been controversial, particularly among academics who fear that by emphasizing matters of efficiency and cost reduction important process values may be lost.[83]

Sampling is uniquely interesting in this context, because although it is a tool of managerial judging in that it helps judges resolve complex discovery disputes in a cost-efficient manner, it does so by providing the judge with new information about the merits of the dispute. A judge who grants a motion to compel after reviewing a sample of the requested discovery, and determining that its "importance to the case" outweighs its cost, is making a merits-based determination, not one based solely on cost or efficiency concerns.[84] As such, sampling responds to some of the concerns of critics of the managerial judging model, who fear that pretrial management is only about reducing costs and settling issues, with little consideration for the merits of the individual case.[85] Sampling allows judges to inquire more fully into the

---

81 See discussion *supra* Part I.A, concerning the specific reference to sampling in the Committee Notes to the 2006 e-discovery amendments. Notably, Judge Scheindlin was a member of the Judicial Conference Advisory Committee on Civil Rules from 1998–2005, and served on the Discovery Subcommittee when the 2006 amendments were drafted.

82 Sec. & Exch. Comm'n v. Collins & Aikman Corp., 256 F.R.D. 403, 418 (S.D.N.Y. 2009).

83 *See* James S. Kakalik et al., *Just, Speedy, and Inexpensive? An Evaluation of Judicial Case Management Under the Civil Justice Reform Act*, 49 ALA. L. REV. 17, 48 (1997); Judith Resnik, *Managerial Judges*, 96 HARV. L. REV. 374, 380 (1982).

84 This is an important counter to a major concern of judicial management opponents—that giving such discretionary power to trial judges without "particularized guidance" is improper and that "neither the drive for efficiency or finality . . . would properly warrant" such unstructured discretion. *See* Steven S. Gensler, *Judicial Case Management: Caught in the Crossfire*, 60 DUKE L.J. 669, 725 (2010).

85 *See* Resnik, *supra* note 83, at 380. Of course, as *McPeek* and *Zubulake* both illustrate,

legal merits of a discovery dispute before ruling on that dispute. As such, it represents an interesting compromise in the debate over managerial judging—a managerialist technique that enables a judge to obtain a neutral, merits-based determination on a disputed legal issue.[86]

This point was well made by Judge Scheindlin in *Zubulake I*, when she noted that sampling would provide her with "tangible evidence of what the backup tapes may have to offer."[87] Both she and Judge Facciola in *McPeek* made it clear that they believed the purpose of sampling was to provide them with additional factual information so they could make better decisions more closely tied to the merits of the disputes.[88]

It is indisputable that the federal courts have moved closer to a managerial judging model in the past thirty years, and that control over discovery has been at the center of that changed role.[89] The move to managerial judging has engendered serious concerns. The first and most important is that judicial decisions regarding scheduling, discovery, and dispute resolution are being made primarily on the basis of cost and efficiency rather than careful consideration of cases' underlying merits. Such concerns are manifested in fears that judges may be overzealously promoting settlement,[90] cutting off promising but costly lines of discovery, imposing unnecessarily stringent time deadlines for

---

sampling also provides the judge with particularized information about the costs of discovery in the individual case.

86 There is, of course, a certain ambiguity in the use of the term "merits" in this discussion. The precise "merits" on which the judge seeks information through sampling are the merits of the discovery dispute, not the merits of the litigation as a whole. In that regard, our endorsement of sampling and some prejudging of the merits should be distinguished from those scholars advocating the use of sampling to *resolve* merit determinations such as findings of liability or damages. *See, e.g.*, Laurens Walker & John Monahan, *Sampling Damages*, 83 IOWA L. REV. 545 (1998). A merits decision in the sampling context would not later be controlling: it is possible for the plaintiff's position on the discovery dispute to be meritorious (if, for example, the discovery provides valuable information about defendant's liability), but for the case itself to lack merit (if, for example, there is no proof of causation). Nonetheless, in most cases there is likely to be a strong positive relationship between the merits of the discovery dispute and the merits of the case.

87 217 F.R.D. 309, 324 (S.D.N.Y. 2003).

88 *Id.*; McPeek v. Ashcroft, 202 F.R.D. 31, 34–35 (2001).

89 Gensler, *supra* note 84, at 671–72 ("Starting in 1983 . . . a series of amendments have enshrined active judicial case management into the Federal Rules of Civil Procedure . . . formally validating it as a favored practice while encouraging and enabling it by giving district judges an ever-expanding set of case-management tools to be used in its pursuit.").

90 *See, e.g.*, E. Donald Elliott, *Managerial Judging and the Evolution of Procedure*, 53 U. CHI. L. REV. 306, 323 (1986); Patrick E. Higginbotham, *Judge Robert A. Ainsworth, Jr. Memorial Lecture, Loyola University School of Law: So Why Do We Call Them Trial Courts?*, 55 SMU L. REV. 1405, 1405 (2002); Peter H. Schuck, *The Role of Judges in Settling Complex Cases: The Agent Orange Example*, 53 U. CHI. L. REV. 337, 347–48 (1986); Elizabeth G. Thornburg, *The Managerial Judge Goes to Trial*, 44 U. RICH. L. REV. 1261, 1264 (2010). *But see* Robert F. Peckham, *The Federal Judge as a Case Manager: The New Role in Guiding a Case from Filing to Disposition*, 69 CALIF. L. REV. 770, 772 (1981); Stephen N. Subrin, *The Limitations of Transsubstantive Procedure: An Essay on Adjusting the "One Size Fits All" Assumption*, 87 DENV. U. L. REV. 377, 389 (2010).

completion of pretrial preparation, and generally making the efficient disposition of cases a greater concern than just adjudication on the merits.[91] While these are undoubtedly dangers, there is serious disagreement among lawyers and judges as to whether they are the inevitable results of the managerial judging model or consequences of that model being applied improperly.[92]

Two other concerns are the unreviewability of most of the decisions made by judges in their managerial capacity, and the bifurcation of the judge's role between a magistrate judge who handles pretrial procedures and a district judge who adjudicates the merits of the case, but it is not clear how great a problem these concerns actually pose.[93] Given the final judgment rule and the deferential standard of review applied to many district court decisions, a large amount of the decision-making done by district court judges is already effectively insulated from appellate review. How serious a problem this is depends in large part on the validity of the first concern—whether pretrial decision-making is systematically skewed away from merits-based considerations. Similarly, the concern over bifurcation of judicial roles seems largely based on similar concerns that magistrate judges will have less regard for the underlying merits of cases.

Sampling is an extension of the managerial judging paradigm in that it gives the judge even greater leeway to shape the pretrial process. While judges already had the power to deny discovery or shift costs under the proportionality standard,[94] sampling gives judges the power to reformulate, not merely limit the discovery request. We believe implicit in the power granted under 26(b)(2)(B) and (C) is the power of the court to order discovery it believes will be useful in determining the issues before it, even if neither party has requested such discovery.[95]

---

91 *See* Paul D. Carrington & Roger C. Cramton, *Judicial Independence in Excess: Reviving the Judicial Duty of the Supreme Court*, 98 CORNELL L. REV. 587, 627–28 (2009); Resnik, *supra* note 83, at 424–25.

92 Professor Resnik warns that even though such extensive pretrial involvement creates a more intimate relationship between the judge and the parties, "neither the Supreme Court, the lower federal courts, nor Congress has considered the effect of judicial management on impartiality" of judges. Resnik, *supra* note 83, at 428; *see also* Robert G. Bone, *Who Decides? A Critical Look at Procedural Discretion*, 28 CARDOZO L. REV. 1961, 1963 (2007); Gensler, *supra* note 84, at 688–743; Jay Tidmarsh, *Pound's Century, and Ours*, 81 NOTRE DAME L. REV. 513, 559 (2006) ("Most case management powers have little or no effect on the metrics of cost, delay, and participant satisfaction.").

93 Interestingly, surveys conducted by the FJC, the ABA, and a joint survey by the IAALS and ACTL all showed "strong support for active judicial case management" from lawyers. *See* Gensler, *supra* note 84, at 687.

94 *See* FED. R. CIV. P. 26(b)(2)(C).

95 The Rule's command that the court must apply the proportionality test to limit discovery "on motion or on its own" is the source of this power. *See* FED. R. CIV. P. 26(b)(2)(C). Accordingly, if the court determines sampling data is required to resolve a discovery issue under Rule 26(b)(2)(C), it is free to order such discovery on its own. Of course, such sampling discovery will usually be a subset of the discovery being sought by one of the parties, but not necessarily. For example, a court might order discovery of a sample of difficult-to-access

Yet sampling is also different from traditional managerial judging, because, when properly understood and applied, it should function as an aid to merits-based adjudication, not an alternative to it. Sampling makes it possible for judges to get a sense of the actual content of the discovery being sought before ruling on its scope, duration and allocation of costs. It therefore has the potential to make merits considerations a stronger, more salient element in decision-making during the pretrial process and to counterbalance the cost and efficiency concerns that critics fear dominate the decision-making process at the pretrial stage. For example, Donald Elliott points out that narrowing issues through managerial judging may have the same impact on litigation as the grant of a motion to dismiss or partial summary judgment. When deciding such motions, a judge "must act according to law and provide a reasoned justification[] subject to appellate review."[96] A judge who forecloses lines of inquiry based on managerial judging, in contrast, will rarely base her decisions "on the legal merits of the parties' positions."[97]

Consider then the judge who forecloses or shifts costs with respect to discovery on a certain line of inquiry after a well-constructed sampling indicates that there is not much chance the plaintiff will be able to support its position with valuable evidence. This is clearly a "merits-based" determination and very likely a dispositive one; if the sampling had revealed a strong likelihood that the discovery sought would have proven valuable, the issue would have been resolved in plaintiff's favor. To be sure, this is not quite a decision "according to law" for which the judge must provide a reasoned justification, but it is a factual determination regarding the merits, based on evidence specific to the case, for which the judge will probably write a reasoned opinion applying the Rule 26(b)(2)(C) proportionality factors. It is unlikely to be subject to appellate review, but is at least potentially and theoretically constrained by the "abuse of discretion" standard.[98] The factual determination involved is, admittedly, a probabilistic one, as any factual determination based on a sampling methodology must be, but such probabilistic approaches to merits issues are certainly not unfamiliar to courts (for example, in preliminary injunction motions).[99]

---

electronic data in order to determine more accurately the per unit cost of producing that data, even if neither side has requested such cost information. Of course, the court's power is still limited to resolving the discovery dispute pending before it. This is not a free-ranging power to investigate the case through its own processes, as exists in some civil law courts. *See, e.g.*, JAMES G. APPLE & ROBERT P. DEYLING, FED. JUDICIAL CTR., A PRIMER ON THE CIVIL LAW SYSTEM 30 (1995).

96 Elliot, *supra* note 90, at 311.

97 *Id.*

98 *See* Thornburg, *supra* note 90, at 1295–96.

99 *See* 11A CHARLES ALAN WRIGHT, ARTHUR R. MILLER & MARY KAY KANE, FEDERAL PRACTICE AND PROCEDURE § 2948.3 (2d ed. 1995).

Like preliminary injunction motions, the application of sampling techniques to discovery disputes also involves a complex balance of merits-based concerns with considerations of cost and expediency. The fact that the preliminary injunction is being sought indicates that there is a perceived need for the court to act on less than a fully developed legal and factual record, yet within those constraining parameters the court strives to make an accurate—if probabilistic—determination of the merits in accordance with the law and the facts. Sampling is similar in that it is also applied when conditions of cost or expediency are said to require immediate judicial intervention,[100] yet the court focuses its inquiry primarily on merits-based considerations, namely the likely content of the information being sought.

We believe that this approach to sampling as a merits-based tool of managerial judging is the best perspective for understanding the sampling technique and developing a coherent and normatively justifiable set of "best practices" for its use. Again, Judge Scheindlin's rulings in the *Zubulake* cases provide a useful illustration. In *Zubulake I* she recognized that application of the complex legal standard for disclosure of relatively inaccessible electronically stored information[101] required more and better factual information than she had concerning the content and cost of the disputed information. Accordingly, she ordered production of a "small sample" of that information, to provide herself with "tangible evidence" on these issues.[102] In *Zubulake III*, she analyzed that sample and concluded that it provided relevant evidence to support plaintiff's claims, primarily in that it contradicted a number of statements defendants had made in opposing those claims. At the same time, she concluded that "*none* of [the sampled e-mails] provide[d] any direct evidence of discrimination."[103] Applying these factual findings to the relevant legal standards, she held that 25% of the cost of restoration of the backup tapes should be shifted to plaintiffs.[104]

The *Zubulake* opinions are clearly an exercise in managerial judging. The Judge not only resolved a discovery dispute, but did so in a way that gave substantial weight to considerations of cost and efficiency. Yet the decision in *Zubulake* is most influenced by the Judge's nuanced view of the factual merits of the motion: the sample that revealed that

---

100 As demonstrated more fully *infra*, sampling is usually invoked after a party moves for a protective order arguing that the cost and burdens of the discovery sought outweigh its likely benefits.

101 In that opinion, Judge Scheindlin actually considers and reformulates a seven-factor version of the proportionality test, which she believes should be applied to the issue of cost-shifting with regard to reasonably inaccessible electronically stored data. That test predates the 2006 e-discovery amendment to Rule 26(b)(2)(B) and may have been partially superseded by it. *Zubulake I*, 217 F.R.D. at 322.

102 *Id.* at 324.

103 *Zubulake III*, 216 F.R.D. at 286.

104 *Id.* at 286, 289.

the disputed information had some but not a great deal of relevance to the plaintiff's claims. As such, it is certainly an individualized decision on the merits.

We believe the *Zubulake* approach can act as a model of good sampling procedure. Note first that Judge Scheindlin treats sampling as "discovery about discovery," a technique that enables her to find out more information about the discovery dispute before her so that she can render a finer-grained, more merits-based ruling on the motion. Implicit in this approach are a number of assumptions:

> First, sampling is primarily for the benefit of the judge, not the parties. Procedurally, this means that a party cannot move for sampling to take place or for a protective order against it. Rather, the decision to use sampling is within the discretion of the judge, to be used when the judge believes it will be helpful in resolving a discovery motion that is pending before her. Of course, there is nothing to prevent a party from suggesting to a judge that sampling might be useful in resolving a pending discovery issue.[105]

The recognition that sampling is for the benefit of the judge also implies a broad general standard for its appropriate use. Sampling should be used to obtain the maximum amount of useful information for the judge at the minimum cost. While this may seem almost tautologous, we will see that it can be quite useful in some cases in deciding between the various sampling methods that have been utilized by different courts. The "maximization of useful information to the judge" standard requires a consideration of the limits of the court in reviewing and analyzing sampling data (even with the help of briefing by the parties) and also implies that the court should have very broad discretion in shaping the scope and amount of sampling that takes place.

This approach also implies rejection of certain other conceptions of sampling. For instance, sampling should not be confused with "phased discovery,"[106] such as partial granting of a motion to compel. While a sampling order may, in some cases, closely resemble that of partial

---

105 *See, e.g.*, Class Plaintiffs' Proposed Litigation Plan at 14 (Aug. 22, 2005) (Exhibit 8 to Class Plaintiffs' Response to the Court's Memorandum of June 6, 2005), Schwab v. Philip Morris USA, Inc., 449 F. Supp. 2d 992 (E.D.N.Y. 2006) (No. CV-0401945) (proposing a plan for sampling of statistical evidence regarding class value in a mass torts action, which was adopted by Judge Weinstein). For the same reason, we reject the suggestion made by some commentators that sampling should be an automatic prerequisite to the resolution of discovery disputes involving backup tapes or other e-discovery. *See* Robert E. Altman & Benjamin Lewis, *Cost-Shifting in ESI Discovery Disputes: A Five Factor Test to Promote Consistency and Set Party Expectations*, 36 N. KY. L. REV. 569, 598–99 (2009).

106 The commentary to Rule 26 describes phased discovery as "regulating the timing and sequence of discovery . . . taking discovery from the most important or the most accessible sources before determining whether there is any need to cast the discovery net more widely." FED. R. CIV. P. 26 cmt.

phased discovery, the purpose of the two orders is quite different. The court uses sampling to resolve the entire discovery dispute between the parties and to allocate the costs on a merits-based application of the proportionality standard. A court orders phased discovery in the hope that the information disclosed in that partial discovery will make further discovery unnecessary.[107] Furthermore, partial phased discovery may sometimes represent a compromise of the discovery motion, giving each party a portion of what they sought. Sampling is a way to resolve the discovery dispute on the merits.

Similarly, sampling should not be confused with a threshold motion, a merits-based hurdle that the party seeking discovery must pass before he becomes entitled to the benefits of the discovery process. Many have argued that the decisions in *Twombly* and *Iqbal* are an attempt to convert Rule 12(b)(6) into such a motion,[108] and some have made similar arguments regarding class certification.[109] While sampling will help ensure that the discovery motion is decided on its merits, the merits of the discovery motion are not the same as the merits of the overall claim. For example, plaintiffs would certainly be entitled to free discovery of data which sampling showed was strongly relevant to their claims of negligence, even if those plaintiffs were not yet able to provide satisfactory proof of causation. Even more fundamentally, however, the concept of "hurdles" implies that the party that seeking discovery has a presumption against him that must be overcome. The proportionality standard, to the contrary, provides for a multi-factor balancing test in which the presumption, if anywhere, is in favor of full disclosure.[110]

The existence of sampling should also not be used as a basis for seeking changes in pleading standards and other aspects of the pretrial litigation process. Given the strong connection evident in *Twombly* and

---

107 Theoretically, the main difference between the two techniques is the judge's ex ante belief about the existence of valuable information in the discovery requested. In partial phased discovery, the judge believes at least some portion of the discovery sought meets the proportionality test, and orders discovery of that portion. In sampling, the judge is uncertain whether any portion of the discovery sought meets the standard, and orders a sample to get better information about the discovery being sought. As a practical matter, the main difference is that phased discovery may include 50% or more of the discovery sought, while, for reasons stated *infra*, we believe sampling should be limited to 15% to 25% of the costs of the total discovery sought.

108 Miller, *supra* note 12, at 22 ("[P]lausibility pleading [has] undone . . . the limited function of the Rule 12(b)(6) motion[] . . . [and] has granted virtually unbridled discretion to district court judges . . . . [creating] a concern that some judges will allow their own views on various substantive matters to intrude on their decisionmaking . . . . ").

109 *See, e.g.*, 1 JOSEPH M. MCLAUGHLIN, MCLAUGHLIN ON CLASS ACTIONS § 3:12 (8th ed. 2011); Judith Resnik, *Fairness in Numbers: A Comment on* AT&T v. Concepcion*,* Wal-Mart v. Dukes*, and* Turner v. Rogers, 125 HARV. L. REV. 78, 149 (2011).

110 *See, e.g.*, Rachel Hytken, *Electronic Discovery: To What Extent Do the 2006 Amendments Satisfy Their Purposes?*, 12 LEWIS & CLARK L. REV. 875, 887 (2008). While it is true that Rule 26(b)(2)(B) imposes a requirement of "good cause" to obtain electronically stored data that is not reasonably accessible, the good cause requirement seems to be the functional equivalent of the 26(b)(2)(C) proportionality test.

Iqbal between heightened pleading standards and fears of abusive and costly discovery, it is tempting for litigants to argue for the application of a relatively low pleading standard based on the assurance that sampling techniques will be applied to prevent abusive discovery. The problem with such arguments is that a judge who denied a motion to dismiss on the grounds that costly or abusive discovery can be avoided at the pretrial stage, would likely feel obligated to apply a higher-than-usual standard of relevance to the proportionality analysis in subsequent discovery disputes.[111] Again, this goes against the principle of neutral merits-based adjudication of discovery disputes that sampling is meant to promote.

A final charge that can be made against the widespread use of sampling is that it necessarily involves the court in prejudging the merits of the underlying dispute. In some respects, this objection is absolutely valid. Indeed, the entire argument thus far has shown that to properly adjudicate discovery disputes in accordance with the proportionality standard, courts must develop information regarding the merits of the underlying dispute, and sampling is generally the optimal method of obtaining such information. The question is not whether this constitutes prejudgment, but whether such prejudgment is harmful to the litigation process, and, if so, whether there are ways this harm can be minimized.

First, it is worth noting that neither the rule-makers nor the courts seem nearly as concerned about prejudgment as they once did. Emblematic of that change is the diminished significance of *Eisen v. Carlisle & Jacquelin*,[112] in which the Supreme Court strongly criticized the use of a merits-based inquiry as the basis for cost-shifting of class action notice fees made "in the absence of established safeguards."[113] In last year's *Wal-Mart* decision,[114] the Supreme Court itself engaged in substantial consideration of the underlying merits in ruling on the class certification motion, and freely admitted it was doing so.[115] In light of the growth of managerial judging obligations, the *Wal-Mart* majority

111 We are speaking here, of course, about arguments tied to an individual case, like the one in *Iqbal*, in which plaintiffs argued that discovery could be limited, an argument specifically rejected in the majority opinion. We do not mean to foreclose the possibility of arguing for a general lowering of the pleading standard (perhaps by Federal Rule) on the grounds that sampling and other managerial judging techniques have significantly reduced the problem of abusive discovery. Unfortunately, the empirical evidence to support such an argument does not yet appear to be available.

112 417 U.S. 156, 178 (1974).

113 *Id.* at 178.

114 Wal-Mart Stores, Inc. v. Dukes, 131 S.Ct. 2541 (2011).

115 Justice Scalia, writing for the majority, acknowledged that "[a] 'rigorous analysis' [of the class certification issues] will entail some overlap with the merits of the plaintiff's underlying claim," but observed that "[t]he necessity of touching aspects of the merits in order to resolve preliminary matters, *e.g.*, jurisdiction and venue, is a familiar feature of litigation." *Id.* at 2551–52. The *Eisen* decision was essentially limited to its facts, and its warnings about the lack of procedural safeguards in pretrial merits determinations were effectively rejected. *Id.* at 2552 n.6.

certainly has a point. A rule against prejudging that asks judges to eschew any consideration of the merits prior to trial would be extremely detrimental to current pretrial practice. Given the rarity of trials, a merits-based decision at the pretrial phase is the only merits-based judgment the parties are likely to obtain. Moreover, given the importance of careful and nuanced considerations of the individualized merits in Rule 12(b)(6) motions, class certification, and discovery matters, a general rule against prejudgments would clearly do more harm than good. Rather, as we have argued in this Section, one of the main benefits of sampling is that it allows for judges to make better prejudgments on the merits.

That said, however, the question remains whether concerns over prejudging can be minimized. Those concerns are largely of two types—lack of trial-level procedural safeguards in pretrial hearings, and a loss of neutrality or impartiality by judges unduly influenced by the pretrial arguments or disclosures. As to the former, one cannot expect every pretrial hearing to have the procedural safeguards of a trial, but the most basic aspects of due process—notice to all parties, an opportunity to be heard, and transparency of decision-making—can and should be preserved in any sampling procedure.

As to the second concern, it is worth noting that in many other prejudging procedures, a degree of tentativeness is recognized and opportunities for revision are maintained. Preliminary injunctions require the winning party to post a bond. Certified classes may be decertified. The fact that the majority of sampling decisions involve cost-shifting seems to us to provide a similar recognition of tentativeness. A plaintiff who is informed that he will have to bear the costs of further discovery has not lost his case. He has not even lost his chance at further discovery. But he has received a powerful signal that the judge does not think much of the evidence he has adduced thus far. Similarly, a judge who shifts part of the costs of discovery, as Judge Scheindlin did in *Zubulake*, may be sending a carefully calibrated signal to the parties as to "where things stand," which may guide them in further pretrial actions or promote a settlement which mirrors the underlying merits of the claim. In short, if judges and parties recognize that a cost-shifting decision is not a final judgment, but a way of evaluating where things currently stand, at least with regard to the particular issue before the court, it is likely to do more good than harm for the litigation process.[116]

II. STATISTICAL AND DESCRIPTIVE ANALYSIS OF REPORTED SAMPLING

116 This does not address the question of whether costs, once shifted, should be shifted back if parties are able to make a stronger showing at a later stage. That and related issues are discussed in Part III, *infra*.

CASES

A. *The Patchwork of Judicial Approaches*

We turn now from our normative justification for sampling to an analysis of the ground-level decision-making about sampling. We have identified forty cases since 1999 that discuss sampling.[117] Although these cases, taken as a whole, demonstrate that sampling has become an established technique for adjudicating electronic discovery disputes, our analysis of the reported cases has also uncovered a wide disparity of approaches in the ways in which courts conceive and implement sampling methods.

At the most basic level, courts differ as to what the purpose of sampling is. Most courts approach sampling as information gathering for purposes of applying the proportionality or "good cause" standards. This is the purpose conceived of in both *McPeek* and *Zubulake,* and the one we adopt here. The purpose in ordering sampling in those cases was not to provide the requesting party with documents (although that certainly was a consequence), but rather to reduce uncertainty in the court's analysis.[118] Some courts, however, have used sampling as a way to incrementalize, or "phase," discovery in order to provide a middle ground or compromise between the respective positions of the parties. For instance, the district court in *Barrera v. Boughton* ordered discovery as part of a "phased approach,"[119] in order to constrain costs. This is an approach, as noted, that we disagree with, but its use as a justification for sampling is nonetheless worth noting.

A second disagreement among the courts is whether the information gathered through sampling should be used to shift the costs of additional discovery or cut-off discovery entirely. Both approaches are arguably consistent with the goal of minimizing undue expenses, as contemplated by the proportionality standard. But, as discussed herein, there are reasons for preferring different results in different situations.[120]

A third major issue in sampling of e-discovery cases is precisely what sort of methodology will be employed. Courts have devised a number of methods for sampling, but neither courts nor commentators

---

117 For a list of these cases, see *infra* note 131. Although the descriptive portions of this Article include considerations of all the reported cases, state and federal, for reasons set forth *infra*, the statistical analysis is limited to thirty-two reported federal court decisions.

118 *See* David Degnan, *Accounting for the Costs of Electronic Discovery*, 12 MINN. J.L. SCI. & TECH. 151, 173–74 (2011) ("Sampling allows the requesting parties to take a snap shot of the producing party's files and draw conclusions of the whole population based on those findings.").

119 No. 3:07cv1436, 2010 WL 3926070, at *3 (D. Conn. Sept. 30, 2010); *see also* Haka v. Lincoln Cnty., 246 F.R.D. 577, 579 (W.D. Wis. 2007) (ordering the parties to proceed "incrementally").

120 *See infra* Part III.D.

have provided a justification for the methodology selected. For instance, the approach adopted in *McPeek* is what we will call the "court order" approach: the court selects which backup tapes will be sampled based on the court's understanding of what tapes are most likely to yield relevant documents.[121] A court may request sampling proposals from the parties, and may even allow the deposition of persons familiar with the data storage system in order to best inform the court's order.[122] A second approach, utilized in *Zubulake*, is to delimit the sample size but leave the selection of tapes to be sampled to the requesting party. This approach will hereafter be referred to as the "best-case scenario" approach, and the theory behind it is that the requesting party will have the most information about what it is seeking and therefore can most effectively select the sample that will yield relevant documents. At least one court has also applied the converse of this approach: allowing the party bearing the potential costs (i.e., the responding party) to define the terms of the sample, subject to court approval.[123] A third more deferential approach—identified here as the "stipulation" approach—is to leave the decision of what to sample to the parties. A fourth approach is scientific sampling. That is, randomized sampling supported by a scientifically sound methodology as likely advocated by an expert witness. As the Fifth Circuit has explained, "[b]y random sampling, we mean adhering to a statistically sound protocol for sampling documents" and the use of expert assistance in constructing any protocol.[124]

Scientific sampling has been suggested by some courts as an alternative to the other three sampling approaches based on criticisms that non-scientific sampling does not provide accurate data upon which cost-shifting analysis may be based.[125] For instance, the Seventh Circuit has expressed concern with sampling procedures that are "inherently arbitrary" and lack a "logical foundation" because the sample is entirely up to one party's discretion and creates "every incentive to cherry-pick,"

---

121 *See, e.g.*, Hagemeyer N. Am., Inc. v. Gateway Data Sci. Corp., 222 F.R.D. 594, 603 (E.D. Wis. 2004) (implementing the methodological approach).

122 *See* Zurich Am. Ins. Co. v. Ace Am. Reinsurance Co., No. 05 Civ. 9170, 2006 WL 3771090, at *2 (S.D.N.Y. Dec. 22, 2006) ("The parties shall therefore propose a protocol for sampling . . . . In order to facilitate that process, counsel may take the deposition of . . . persons familiar with [defendant's] data storage system.").

123 Ingersoll v. Farmland Foods, Inc., No. 10–6046, 2011 WL 1131129 (W.D. Mo. Mar. 28, 2011).

124 *In re* Vioxx Prods. Liab. Litig., Nos. 06-30378, 06-30379, 2006 WL 1726675, at *2 n.5 (5th Cir. May 26, 2006) (per curiam); *see also* Degnan, *supra* note 118, at 174 ("[T]he sample must be random, must compare the same type of variables, must have a representative sample size, and must use a statistically valid method that is planned beforehand.").

125 In other cases, courts have simply crafted the sample in such a way to avoid representativeness concerns. *See, e.g.*, Quinby v. WestLB AG, 245 F.R.D. 94, 107 (S.D.N.Y. 2006) (requiring the sampling of certain e-mails in response to the defendant's objection that sampling is inherently flawed because there is a stronger likelihood of responsive documents from the sampled period than from other periods).

thereby producing an unrepresentative sample of the universe of documents.[126] In fact, at least one commentator has argued that the absence of a scientifically valid sampling approach could pose due process issues:

> A sample cannot be extrapolated if [it] is not statistically valid, because the margin of error would not produce results that are accurate with a high degree of certainty. A court will likely overturn any such sampling protocol on due process grounds if the margin of error is too high. . . . '[A court] may adopt statistical sampling and extrapolation as a case management tool only when the specific methodology to be used is tailored to produce a result at least as fair and accurate as would be produced by traditional particularistic fact-finding methods.'[127]

Related to this disagreement over what sampling methodology to use, there has also been wide variation among courts in terms of how large of a sample should be selected. Some courts have sampled as much as a quarter of the backup tape documents being requested. Others have opted for a smaller sample, limiting it to a few backup tapes.

A fourth point of divergence is the variable factors that cause a court to order or deny sampling, or to ultimately shift costs after the sampling has occurred. Court decisions are frequently unclear as to which factors lead them to order or reject sampling. In most cases, sampling is ordered without discussion of such reasons. The next Section seeks to apply statistical methods to delineate the factors that may influence a court's decision.

One final issue raised by the Sedona Conference, but which has received limited discussion from courts, is whether the responding party should always pay for the sampling.[128] Courts have acted consistently in this regard: almost universally courts have imposed the cost of sampling on the producing party. Yet no reported decision provides any justification for such an allocation. The closest a court has come is in *Kipperman v. Onex Corp.*[129] where the court made the requesting party "the guarantor of the search's success" and granted the producing party the right to demand fees if the sampling produced little discoverable

---

126 Am. Nat'l Bank & Trust Co. of Chi. v. Equitable Life Assurance Soc'y, 406 F.3d 867, 879 (7th Cir. 2005). Courts have emphasized similar methodological concerns in other aspects of e-discovery. *See* William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134 (S.D.N.Y. 2009) (discussing keyword searches); Gonzales v. Google, Inc., 234 F.R.D. 674 (N.D. Cal. 2006) (discussing URLs); Herbert L. Roitblat, *Search and Information Retrieval Science*, 8 SEDONA CONF. J. 225 (2007).

127 *See, e.g.*, Degnan, *supra* note 118, at 174–75 (quoting Scottsdale Mem. Health v. Maricopa Cnty., 228 P.3d 117, 134 (Ariz. Ct. App. 2010)).

128 Laura E. Ellsworth & Robert Pass, *Cost Shifting in Electronic Discovery*, 5 SEDONA CONF. J. 125, 147 (2004).

129 260 F.R.D. 682, 690 (N.D. Ga. 2009).

information.[130]

B. *How Courts Sample*

Given the wide qualitative variations in when and how courts sample in electronic discovery disputes, we have undertaken an empirical investigation to make sense of this patchwork of approaches to discovery sampling.

1. Data Description and Methodology

For our empirical models, we rely on data from federal district court cases from 2001 to the present. Our database includes thirty-two federal district court opinions contemplating the use of sampling techniques to manage electronic discovery cost disputes.[131] We thus

---

130 Ultimately, in *Kipperman*, the sampling was a success and so the court never revisited the issue of shifting the costs for the initial sample. Additionally, in *Makrakis v. Demelis*, No. 09-706-C, 2010 WL 3004337, at *2 (Mass. Super. Ct. July 13, 2010), the court ordered that the requesting party bear the initial cost of the sampling with leave to seek reimbursement for those costs if the sampling proved successful. However, since *Makrakis* was a state court decision, it has not been included in our data set.

131 Our data set is drawn from the following cases: Ingersoll v. Farmland Foods, Inc., No. 10-6046, 2011 WL 1131129 (W.D. Mo. Mar. 28, 2011); Barrera v. Boughton, No. 3:07cv1436, 2010 WL 3926070, at *3 (D. Conn. Sept. 30, 2010); *Kipperman*, 260 F.R.D. 682; D'Onofrio v. SFX Sports Grp., Inc., 256 F.R.D. 277 (D.D.C. 2009); Sec. & Exch. Comm'n v. Collins & Aikman Corp., 256 F.R.D. 403 (S.D.N.Y. 2009); *William A. Gross Constr. Assocs.*, 256 F.R.D. 134; Major Tours, Inc. v. Colorel, No. 05-3091, 2009 WL 3446761, at *6, n.7 (D.N.J. Oct. 20, 2009); U & I Corp. v. Advanced Med. Design, Inc., 251 F.R.D. 667 (M.D. Fla. 2008); Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251 (D. Md. 2008); Parkdale Am., LLC v. Travelers Cas. & Sur. Co. of Am., Inc., No. 3:06CV78, 2007 WL 4165247 (W.D.N.C. Nov. 19, 2007); Haka v. Lincoln Cnty., 246 F.R.D. 577 (W.D. Wis. 2007); Hill v. Eddie Bauer, 242 F.R.D. 556 (C.D. Cal. 2007); Zurich Am. Ins. Co. v. Ace Am. Reinsurance Co., No. 05 Civ. 9170, 2006 WL 3771090 (S.D.N.Y. Dec. 22, 2006); Quinby v. WestLB AG, 245 F.R.D. 94 (S.D.N.Y. 2006); Semsroth v. City of Wichita, 239 F.R.D. 630 (D. Kan. 2006); Thompson v. Jiffy Lube Intern., Inc., No. 05-1203, 2006 WL 1174040 (D. Kan. May 1, 2006); Hopson v. Mayor & City Council of Balt., 232 F.R.D. 228 (D. Md. 2005); Wiginton v. CB Richard Ellis, Inc., 229 F.R.D. 568 (N.D. Ill. 2004); Hagemeyer N. Am., Inc. v. Gateway Data Scis. Corp., 222 F.R.D. 594 (E.D. Wis. 2004); Medtronic Sofamor Danek, Inc. v. Michelson, 229 F.R.D. 550 (W.D. Tenn. 2003); *Zubulake III*, 216 F.R.D. 280 (S.D.N.Y. 2003); *Zubulake I*, 217 F.R.D. 309 (S.D.N.Y. 2003); Rowe Entm't, Inc. v. William Morris Agency, Inc., 205 F.R.D. 421 (S.D.N.Y. 2002); Murphy Oil USA, Inc. v. Fluor Daniel, Inc., No. 99-3564, 2002 WL 246439 (E.D. La. Feb. 19, 2002); McPeek v. Ashcroft, 202 F.R.D. 31 (D.D.C. 2001); AAB Joint Venture v. United States, 75 Fed. Cl. 432, 443–44 (Fed. Cl. 2007); Farmers Ins. Co. v. Peterson, 81 P.3d 659 (Okla. 2003); Toshiba Am. Elec. Components, Inc. v. Super. Ct., 21 Cal. Rptr. 3d 532, 541 (Cal. Ct. App. 2004); Makrakis v. Demelis, No. 09-706, 2010 WL 3004337 (Mass. Super. Ct. 2010); Linnen v. A.H. Robins Co., Inc., No. 97-2307, 1999 WL 462015 (Mass. Super. Ct. 1999); Delta Fin. Corp. v. Morrison, 819 N.Y.S.2d 908 (Sup. Ct. 2006); Bank of Am. Corp. v. SR Intern. Bus. Ins. Co., No. 05-CVS-5564, 2006 WL 3093174, at *15 (N.C. Super. Ct. Nov. 1, 2006); Brokaw v. Davol, Inc., Nos. 07-5058, 07-4048, 07-1076, 2011 WL 579039, at n.7 (R.I. Super. Ct. Feb. 15, 2011); Defendant's Response to Plaintiff's Motion to Compel, Calixto v. Watson Bowman Acme Corp., No. 007CV60077 (S.D. Fla. Jan. 26, 2009), 2009 WL 826089; Defendant's Motion for Protective Order, Wyckoff v. United

captured litigation during a period of approximately ten years. The earliest opinion concerning sampling uncovered in the course of our investigation was a 1999 opinion from a state court. We excluded that opinion and other state court opinions from our data set to remove jurisdiction-specific discovery rule variation from our analysis. We searched in the WESTLAW federal "AllCases" database for opinions discussing "sampling," "phased discovery," or a "test-run" in electronic discovery disputes. These search results were checked against a search of the LEXIS "Mega" database using the same search terms. We also searched civil trial court filings through WESTLAW and LEXIS using the same search terms to identify cases in which sampling was proposed but was not necessarily mentioned in a court's opinion. Additionally, we compiled a list of citations both cited in these opinions and citing these opinions, and included any additional cases discussing sampling. For instance, as *McPeek* and *Zubulake* are arguably the originating cases for the sampling methodology, we Shepardized these cases to identify decisions not picked up through keyword searches. Once cases had been identified by this method, we attempted to gather additional information about case characteristics from docket entries using the LEXIS *Total Litigator CourtLink* and Public Access to Court Electronic Records (PACER).

Our multivariate analysis endeavors to shed light on a few central issues relating to sampling: 1) under what circumstances do courts order sampling?; 2) when sampling is ordered, by what method is it conducted?; and 3) does the decision to order sampling play any role in whether cost-shifting or a cut-off of discovery is ordered? Our general, intuitive thesis is that courts make decisions about when and how to sample and whether to shift costs following sampling based on a number of variable factors. Accordingly, in answering these questions, our analysis considers the following independent variables:

1) *Amount in Controversy (AmtCont)*: The logged total amount in controversy as presented to a court by the plaintiff, often in the prayer for relief section of a complaint or in a subsequent filing.

States, No. 307CV02301 (N.D. Cal. Jan. 21, 2009), 2009 WL 3661465; Defendant's Reply in Support of its Motion for Protective Order as to Re-Creation of Back-Up Server Tapes, Mount Sinai Med. Center of Fla., Inc. v. Mckesson Med. Mgmt., LLC, No. 06-21518-Civ-Atlonaga/Turnoff (S.D. Fla. Jan. 22, 2007), 2007 WL 617088; Opposition of Plaintiffs to Defendants' Second Motion to Compel Discovery, *In re* Pharm. Indus. Average Wholesale Price Litig., No. 01-CV-12257 (D. Mass. Feb. 14, 2006), 2006 WL 4578352; Plaintiff's Memorandum Opposing Defendant's Motion for Protective Order, Bergerson v. Deephaven Capital Mgmt., LLC, No. 03-1090 (D. Minn. Oct. 19, 2004), 2004 WL 3370915; Memorandum of Law in Support of Protective Order and Cost-Shifting, Wolf Concept S.A.R.L. v. Eber Bros. Wine & Liquor Corp., No. 607CV06233 (W.D.N.Y. Oct. 16, 2009), 2009 WL 4839819; Memorandum in Opposition to Defendants' Motion for Protective Order, USA, *ex rel.* Jeffrey Judd, Relator, v. Maloy, No. 3-03-241 (S.D. Ohio. Mar. 8, 2006), 2006 WL 1111924; Brief in Support of Defendants' Motion for Protective Order, Camesi v. Univ. of Pittsburgh Med. Ctr., No. 09-CV-085 (W.D. Pa. Aug. 25, 2010), 2010 WL 4545724.

2) *Total Cost of Discovery (CostDisc)*: The logged total cost of ordering the discovery that is at issue in a controversy. To be clear, this variable does *not* reflect the overall cost of discovery in litigation; rather, it reflects the total cost that would be incurred by the producing party in the event that cost-shifting or a limitation of discovery was denied. In most cases this figure incorporates the costs of restoring backup tapes, expert fees, and attorneys' fees associated with reviewing the documents for privilege.

3) *Sampling Cost (CostSamp)*: The logged estimated or actual cost of a sample to be ordered by a court. Where both figures are available, our model uses the estimated costs because the ordering of sampling is an ex ante decision that is made without the benefit of the actual cost data. Where a cost range is presented for the estimated cost of sampling, we have used the high end of the cost range.

4) *Percent of Amount in Controversy (PerAmtCont)*: This variable is calculated by dividing *CostDisc* by *AmtCont* in order to determine the percentage of a potential recovery that is made up by the cost of discovery. This variable functionally captures the third element of the *Zubulake* cost-shifting test: the total cost of production compared to the amount in controversy.[132] This is an important consideration because, as Judge Scheindlin has stated, "where the cost of a sample restoration is significant compared to the value of the suit, or where the suit itself is patently frivolous, even this minor effort *may* be inappropriate."[133]

5) *Percent of Total Cost (PerCostDisc)*: This variable is calculated by dividing *CostSamp* by *CostDisc* to determine the proportional share of the total cost of discovery that is being selected for sampling by a court.

6) *Cost-Shifting or Discovery Cut-off (CostShift)*: This binary variable reflects whether the dispute centers on a request by a producing party that costs be shifted or that certain discovery be cut-off.

7) *Asymmetric Discovery (AsymDisc)*: This binary variable indicates whether the parties had asymmetric discovery obligations. Research indicates that asymmetrical discovery obligations affect parties' abilities to reach agreements in discovery disputes and their settlement decisions.[134] In cases where discovery burdens are symmetrical, parties may be more willing to reach private agreements to limit costs. This variable also, to some extent, reflects the fifth element of the *Zubulake* cost-shifting test: the relative incentives of each party to control costs.[135]

8) *Government as a Party (Govt)*: This binary variable reflects whether the government is a party. While many suits involving the government have asymmetric discovery burdens (which would be

---

132 *Zubulake I*, 217 F.R.D. at 322.

133 *Id.* at 324 n.77.

134 *See* Hay, *supra* note 5, at 30–34 (1995).

135 *Zubulake I*, 217 F.R.D. at 322.

captured by the *AsymDisc* variable),[136] this variable is included to reflect the unique factors involved when a government is a party.

Our methodological choices have been influenced by the size and in some cases incompleteness of our data set. Specifically, there are a sizable number of missing cases for some independent variables. In many cases the cost-related variables (e.g., total cost of discovery and sampling costs) were not reported in the court opinions or parties' pleadings. In those instances, inclusion of those cases in a multivariate analysis will present significant methodological difficulties.[137] Further, even where the data is complete, because many of the variables in our analysis are binary (e.g., whether the court sampled or not), and our overall data set is small, the data is not well suited for multivariate regression analysis. In an attempt to best utilize what data does exist about sampling, our empirical analysis uses two separate statistical approaches. One is a univariate approach involving linear regressions to predict the decision to sample or shift costs. Univariate analyses permit evaluation of the *entire* effect of a particular variable on the decision to sample or shift costs.[138] Linear regression modeling allows us to determine the linear function that most accurately characterizes the relationship between two variables.[139] We have used univariate linear regressions to analyze the effect of the cost variables (amount in controversy, total cost of discovery, and cost of sampling) on the decision to sample or shift costs. In such analyses, the dependent variable is a dichotomy: in our first set of regressions the value of 1 if sampling is ordered and 0 if it is not. The probability of ordering sampling is a nonlinear function of the exogenous variables (described herein). Examining the coefficient estimates and *p*-values allows the identification of the variables that produce statistically significant effects on the decision to sample. In the second set of regressions, the value of 1 is cost-shifting or the cutting off of discovery, and 0 is the authorization of full discovery. Again, the probability of ordering cost-shifting or the cut-off of discovery is a nonlinear function, and the coefficients and *p*-

136 Generally, either the government is being sued and bears the brunt of the discovery burdens, or the government is suing and it is the defendant that bears the costs. In both cases asymmetries often emerge.

137 Specifically, we would either need to exclude those cases from our data set (which would reduce our set to fifteen cases) or, in lieu of deletion, include estimated "dummy" variables for the missing data, which might produce substantially different results. *See* JACOB COHEN ET AL., APPLIED MULTIPLE REGRESSION/CORRELATION ANALYSIS FOR THE BEHAVIORAL SCIENCES 450 (3d ed. 2003) (noting that dropping cases with missing data is disfavored).

138 By contrast, a multivariate comparison shows the aggregate effect of all variables and only details the marginal effect of each variable. While such an analysis could be illuminating, the data set here is not well suited to such a methodology for the reasons described herein.

139 This function is expressed by the equation $y = bx + c$, where y is the predicted value of a dependent variable (here, whether sampling or cost-shifting is ordered), x is the value of the exogenous variable (here, for instance, the cost of discovery), b is the slope of the line of best fit, and c is the y-intercept of the line.

values allow the identification of those variables that have a significant effect on the decision to cut-off discovery or shift costs. We have used this method of analysis for all non-binary independent variables (AmtCont, CostDisc, CostSamp, PerAmtCont, and PerCostDisc).

Our second statistical approach uses two-way classification tables to analyze the interaction between binary variables (for instance, the effect of asymmetric discovery obligations on the decision to sample). We offer our results in two formats: (i) interpretation of the data displayed in each table, and (ii) inferential analysis. In interpreting the tables, the data should be read row-by-row from left to right. The independent variables (e.g., whether discovery obligations are asymmetrical), which only have two possible variations, are displayed in each row. The dependent variables (e.g., whether sampling is ordered) are notated by column. As the independent variable changes (the first row compared to the second), theoretically we should be able to detect whether and how the dependent variable changes based on observed data frequencies. We have used a Fisher's Exact Test, which indicates whether the distribution in the two-by-two matrix is due to something other than chance. If the calculated *p*-value is less than 0.05,[140] the null hypothesis is refuted and the research hypothesis is corroborated (i.e., the distribution is not dependent on chance). The chi-squared test is the most frequently used method of inferential statistical testing for two-by-two contingency tables. Here, however, it is inappropriate because the expected frequencies in any cell are around or less than five.[141] Fisher's Exact Test is most appropriate where the total sample size is small to moderate.[142] Further, because the chi-squared test relies on large sample approximations, in a small sample (as it is here) it could result in skewed approximations, yielding higher calculated values and making it easier to reject incorrectly the null hypothesis. Fisher's Exact Test, on the other hand, calculates a true level of significance.[143] Notably, though, a limitation of this test is that it cannot measure the strength of the effect in the event it is significant.

Before reviewing the results of these analyses, we pause to mention a few shortcomings in our methodology. First, our initial data set—comprising fewer than fifty cases—is a very small sample size. Second, research related to court orders on discovery can raise a number of issues regarding representativeness given the fact that many discovery orders are issued by magistrate judges—sometimes orally—and are

140 The level of statistical significance adopted for testing all hypotheses in this study is 0.05 (the standard measure). This means that there is a five percent chance of being wrong in finding that dependent and independent variables are related when random chance could be the reason for their apparent relationship.

141 THEODORE COLTON, STATISTICS IN MEDICINE 164–65 (1974).

142 HUBERT M. BLALOCK, JR., SOCIAL STATISTICS 292 (2d ed. 1979).

143 C. Frank Starmer et al., *Some Reasons for Not Using the Yates Continuity Correction on 2 x 2 Contingency Tables*, 69 J. AM. STAT. ASS'N 346, 376–78 (1974).

often unreported by publishing services. Even in the cases where a judge's decision ordering sampling is published, there is a likelihood that the parties will independently resolve the issue, either by working out the discovery dispute on their own or settling the case. In particular, our methodology likely undercounts cases in which sampling was denied, as courts may not mention a party's suggestion that the court order sampling.[144] Lastly, on a more theoretical level (a subject discussed at length *infra*), we have ascribed methodological categories to courts' sampling decisions when in all likelihood these courts never intended to design a sampling protocol. A close reading of the cases reveals ad hoc decision-making without any substantial discussion of prior courts' methodological choices. Nonetheless, despite these shortcomings our hope is that through the patterns revealed by the data, and our more general theoretical discussion and guidance for administering sampling, we can provide some coherence to disparate sampling approaches.

2. Results

We have analyzed the data to answer three primary questions: (a) when do courts sample; (b) how is sampling conducted; and (c) in what cases are costs then shifted? Our findings relating to those questions along with other observations relating to the data are set forth herein.

a. When Do Courts Sample?

Our review of district court opinions indicates that in those cases in which sampling was discussed, it was ordered 84.38% of the time. This figure is unsurprising where courts have raised the issue of sampling; intuitively it is unlikely that a court would raise the possibility only to shoot it down. In fact, it appears that in all but one case in which sampling was *not* ordered, it was a party and not the court that suggested sampling. A review of the characteristics of the cases in which sampling was conducted reveals an observed preference for cases in which discovery obligations were asymmetrical (see Table 1). The odds ratios displayed in Table 1 demonstrate that in those cases in which sampling is ordered, there is a 56.25% chance that the parties to the litigation had asymmetric discovery obligations. However, while anecdotally interesting, the presence of asymmetric discovery obligations is not predictive of sampling being ordered, as the results are

144 To account for this concern, we searched district court filings for references to sampling in an effort to capture those cases in which sampling was suggested and summarily dismissed by the court. Nonetheless, not all trial court filings are searchable.

statistically insignificant and the null hypothesis cannot be rejected.[145]

[145] Whether taking the Fisher's Exact Test one- or two-tailed results, the *p*-values are wholly not significant (0.572, one-tailed, 1.00, two-tailed).

Table 1
Asymmetric Discovery Obligations

	Sampling Ordered	Sampling Not Ordered
Asymmetric Discovery Burdens	18 Cases (56.25%)	3 Cases (9.38%)
Equal Burden on Parties	9 Cases (28.13%)	2 Cases (6.25%)

Statistically, similar results are revealed when reviewing the presence of the government as a party. In 25% of the cases in our data set the government was a party. A Fisher's Exact Test reveals that the correlation between the government as a party and sampling being ordered is not significant.[146] This result is not surprising; especially because in 59.38% of the cases in which sampling was ordered, the government was not a party. Nonetheless, the correlation between the government as a party and sampling should not necessarily be rejected. While more often than not sampling is ordered in cases in which the government is not involved, in each of the cases in our data set where the government was a party, the government was the responding party to document requests. As the government is not party to roughly a third of all federal civil cases, this suggests that courts may be more interested in relying on sampling where the government is involved. Such a conclusion is further supported by Judge Francis's commentary in *McPeek* that traditional cost-shifting analyses ignore the fact that a government has to have its employees do the restoration or risk confidential information being seen by someone not employed by the government, and therefore a decision to order additional discovery will necessitate the diversion of government resources.[147] Accordingly, sampling in those cases may suggest a sensitivity for such concerns.

146 Table 2 indicates that the *p*-value for a one-tailed test is 0.506, and again 1.00 for a two-tailed test. Neither of these results is close to significant statistically.

147 McPeek v. Ashcroft, 202 F.R.D. 31, 34 (D.D.C. 2001).

Table 2
Government As Party

	Sampling Ordered	Sampling Not Ordered
Government is a Party	8 Cases (25%)	2 Cases (6.25%)
Government is Not a Party	19 Cases (59.38%)	3 Cases (9.38%)

Arguably the best predictors of the decision to sample—to the extent any exogenous variable can predict a court's decision to sample—are the variables relating to economics of the litigation: the total amount in controversy, the total cost of discovery, and the cost of sampling. A series of linear regressions (the results of which are displayed in Table 3) reveal that the total amount in controversy and the total cost of discovery have a significant correlation with the decision to sample. That is, as the total amount in controversy and cost of discovery rise, it is more likely that a court will order sampling. These results are significant at the $p > 0.05$ level.

Table 3
Monetary Factors

	Coefficient	*P* > |t|
Amount in Controversy (AmtCont)	.2198469	0.040
Total Cost of Discovery (CostDisc)	.2484825	0.026
Cost of Sample (CostSamp)	.1336378	0.417
Percent of Total Amount in Controversy (PerAmtCont)	.3820598	0.739
Percent of Discovery Costs (PerCostDisc)	-1.24611	0.000

However, the cost of a sample has an insignificant effect on the decision to sample. Such a finding is intuitive because the cost of the sample is within a court's control (i.e., if the cost of the sample is too high, the court can decrease the size of the sample). Our findings as to the PerCostDisc variable (cost of sample/total cost of discovery) further confirm this finding; sampling was ordered in all instances in which the cost of the sample was less than or equal to 25% of the total cost of discovery. Of course, this poses a causality question: does the comparatively lower sampling cost compared to the total cost of discovery influence a decision to order sampling? Or is it that courts order sampling and then, in order to effectuate the purposes of sampling, limit the size of the sample? More simply, does the PerCostDisc variable reflect an ex ante or ex post calculation by the court? Because courts can modulate the size of a sample *after* making the decision to sample, it is unlikely that this variable is predictive in most cases. That said, there are two reasons not to dismiss PerCostDisc as having *no* predictive value. First, the cost of sampling compared to the total cost of discovery may be relevant in those cases in which the overall cost of discovery is low enough that sampling would not serve a significant cost-saving purpose. For instance, in *Semsroth v. City of Wichita*,[148] where the expected cost of discovery was a mere $3,374.95, the court rejected sampling because it would not substantially reduce the cost of discovery. The cost of contested discovery was just *too small* to justify sampling. Second, in many cases courts have selected a sample not based on cost but on size. For instance, a court might order production of a fifth of the backup tapes, not the production of as many tapes as can be "bought" for a fifth of the discovery costs.[149] In other words, the value of the *PerCostDisc* metric may be that it indicates a judicial hesitation to order sampling where it would not represent a sufficient cost savings. In any event, there is a significant correlation between that metric and the decision to sample.

We also examined whether the type of cost-minimization analysis being conducted had any effect on the decision to sample (i.e., does cost-shifting versus cut-off-of-discovery analysis make a difference?). There is, however, no identifiable correlation between the type of analysis and the decision to sample. In sum, it appears that the total amount in controversy, the total cost of discovery, and arguably the

148 239 F.R.D. 630, 638–40 (D. Kan. 2006).

149 There is reason to believe that the marginal cost of producing backup tapes decreases with each increase in the number of tapes that are produced, as there are fixed costs associated with production that will be imposed irrespective of whether one or one hundred tapes are produced. Accordingly, a court's decision to order the sampling of a fifth of relevant backup tapes does not mean the court necessarily believes it is ordering an imposition of a fifth of the costs of total production.

presence of the government as a party (although this conclusion isn't necessarily borne out by the data) influence a court's decision to sample.

b. How Courts Sample

In most cases in which sampling is considered by a court, a "best case scenario" (as utilized in *Zubulake*) or court-order methodology (as used in *McPeek*) is employed. Specifically, as indicated in Table 4, in 31.25% of cases the court adopted a "best case scenario" method, and in 37.5% of cases a court-order method was used.

Table 4
Methodology Used

Method	# of Cases
Best Case Scenario Approach	10 Cases (31.25%)
Court Order Approach	12 Cases (37.5%)
Stipulation Approach	7 Cases (21.88%)
Scientific Approach	3 Cases (9.38%)

However, the data reveals a clear division in how courts sample, and courts have provided little if any justification for why they have adopted the sampling method being utilized. For instance, we found no discussion of whether an approach letting the requesting party select the sample is preferable to one in which the court selects the sample. Moreover, despite the fact that in many cases the purpose of sampling is to inform a subsequent cost-shifting analysis by a court, no court has considered whether a court-ordered sample is preferable to a sample selected by agreement of the parties. While courts discussing a scientific methodology have provided a more substantive gloss on their approaches, these discussions usually only appear in the form of criticizing ad hoc approaches to sampling without providing any complete scientific framework for sampling.[150]

A series of statistical analyses of this data revealed no statistically significant results; that is, no readily identifiable correlation between any of the aforementioned independent variables and a court's choice of

150 *See In re* Vioxx Prods. Liab. Litig., Nos. 06-30378, 06-30379, 2006 WL 1726675, at *2 n.5 (5th Cir. May 26, 2006) ("By random sampling, we mean adhering to a statistically sound protocol for sampling documents . . . . The parties must provide expert assistance to the district court in constructing any protocol."); Am. Nat'l Bank & Trust Co. of Chi. v. Equitable Life Assurance Soc'y, 406 F.3d 867, 879 (7th Cir. 2005) (per curiam).

methodology. This result is unsurprising. As our descriptive analysis of the cases indicates, the choice of sampling methodology appears to be entirely case-by-case, and often without consideration of the methodological strengths of any particular sampling approach. This point was put well by the Supreme Court of Oklahoma:

> The parties . . . do not discuss either the varying methods of selecting particular files for sampling or the proper discoverable information in those files that would necessarily support the litigation objective of [p]laintiffs. They may, or may not, agree to a particular method. No authority is cited in support of placing a burden on the trial court to create, *sua sponte*, a statistical sampling discovery technique for parties.[151]

In short, very little guidance is available from the courts as to how to administer sampling and what sort of methodology to use. The data confirm such a characterization.

c. In Which Cases Are Costs Then Shifted?

After courts sample, the data suggests that they are more likely to shift costs or cut-off discovery. Specifically, in 55.55% of cases in which sampling was contemplated, costs were later shifted. These results, displayed in Table 5, are significant at the $p > 0.05$ level.[152]

Table 5
When Are Costs Then Shifted?

	Costs Shifted	**Costs Not Shifted**
Sampling Ordered	10 Cases (55.55%)	4 Cases (22.22%)
Sampling Not Ordered	0 Cases	4 Cases (22.22%)

The results in Table 5 suggest that sampling is ordered in cases in which the court has doubts as to the importance of the information sought through extended electronic discovery. Presumably, the high percentage of costs shifted (55.55%) represents cases where the sample showed little if any relevance to the facts of the case.

---

151 Farmers Ins. Co. v. Peterson, 81 P.3d 659, 661 (Okla. 2003).

152 Using a Fisher's Exact Test these results are significant ($p > 0.023$, one-tailed, 0.023, two-tailed). Notably, though, because of some incompleteness in the data we only have observations for eighteen cases.

The only other variables that have a significant effect on the decision to cut-off discovery or shift costs are the total cost of discovery and the percent of the amount of controversy that is equal to the total cost discovery. Those results are displayed in Table 6.

Table 6
Effect of Costs on Cost-Shifting Analysis

	Coefficient	***P*** **> |*t*|**
Amount in Controversy (AmtCont)	.0539435	0.712
Total Cost of Discovery (CostDisc)	.2912923	0.043
Percent of Total Amount in Controversy (PerAmtCont)	2.387332	0.085

The presence of the government as a party,[153] asymmetric discovery obligations,[154] the total amount in controversy,[155] or whether the analysis was to shift costs or cut-off discovery, had insignificant effects on the decision to cut-off discovery or shift costs.[156] These results are consistent with our description of cost-shifting analysis articulated herein: courts appear to examine the results of sampling along with the total cost of discovery and the percent of the total amount in controversy (in addition to other factors) in deciding whether to shift costs.

III. NORMATIVE IMPLICATIONS OF THE ANALYSIS: RECOMMENDED "BEST PRACTICES" FOR SAMPLING

A. *The Decision to Sample*

The discovery rules themselves, as well as our prior discussion in Part I, indicate that the decision to sample should operate as a function of two sets of factors: those relating to costs and those involving

153 *P* > 0.201, one-tailed, 0.321, two-tailed.
154 *P* > 0.563, one-tailed, 1.000, two-tailed.
155 *P* > 0.712 (see Table 6).
156 *P* > 0.352, one-tailed, 0.630, two-tailed.

uncertainty regarding the relevance and importance of the information sought. The proportionality standard prescribes an explicit balancing test in which the expenses or burdens of proposed discovery are weighed against its likely benefit. Sampling is appropriate when that technique is likely to provide the court with more information on one or both of those elements in a cost-effective manner.

Our empirical study has shown that the total amount in controversy (i.e., the size of the litigation) and the total cost of discovery have a significant correlation with the decision to sample. That is, as the total amount in controversy and cost of discovery rise, it is more likely that a court will order sampling. These results are significant at the $p > 0.05$ level. These findings make intuitive sense. Litigation needs to be of a great enough size and scope so as to justify relatively expansive discovery. When the amount at stake is small, plaintiffs have relatively little incentive to engage in extensive discovery, and defendants have strong arguments for resisting any such requests. There is an even stronger intuitive relationship between the cost of discovery, the ratio of discovery costs to total amount in controversy, and the decision to sample. If discovery costs are low, or the ratio of discovery to amount in controversy is low, the balance between cost and benefit is likely to favor even marginally relevant discovery. In such circumstances the potential cost savings offered by the sampling technique is unnecessary.

Conversely, when discovery costs are high, whether in an absolute sense or as a ratio to the total amount in controversy, the balancing required by the proportionality test will be harder to perform, since it will be necessary to determine not merely whether the information is relevant, but whether it is sufficiently important to the litigation to justify the substantial added costs. Sampling, as we have seen, can frequently provide such finely nuanced insight into the information being sought. Moreover, when costs of discovery are high, in either an absolute or relative sense, the utility of the sampling technique increases because it provides information at a relatively lower cost. *McPeek* and *Zubulake* were both cases where the costs of the requested discovery were substantial enough to justify a preliminary inquiry or fact-finding into the requested discovery.

It seems unwise to prescribe any strict cut-off in the size of the litigation or in the absolute or relative cost of discovery necessary to justify sampling.[157] Given the wide variety of cases that may be presented to a court, and the frequent difficulty of determining the true amount at stake in various kinds of litigation, this is best left to the individual judge. Nonetheless, courts should be aware of cases like

[157] A lower threshold for a sampling procedure may also be appropriate when the discovery is sought from a third party. *See, e.g.*, *In re* Coordinated Pretrial Proceedings in Petroleum Prods. Antitrust Litig., 669 F.2d 620, 623–24 (10th Cir. 1982).

Semsroth[158] and *Parkdale*,[159] where the discovery issues were resolved without sampling and the cost of the discovery sought was $3,374.95 and less than $20,000 respectively. In contrast, the discovery sought in *Zubulake* was estimated at $175,000, which was 1.35% of plaintiff's maximum compensatory recovery of $13 million.[160] Certainly, cases with discovery disputes in the *Zubulake* cost range or higher would be appropriate candidates for the sampling technique.

A second relevant consideration relating to cost is the degree of certainty with which the costs of discovery can be determined. In cases where the actual costs of the discovery are substantially disputed or are otherwise uncertain,[161] sampling provides an effective method for providing a more reliable cost estimate. This may sometimes be necessary to determine if the threshold requirement of Rule 26(b)(2)(B) of "undue burden or cost" is met, or to more effectively apply the proportionality test of Rule 26(b)(2)(C). The *McPeek* case provides an example of a court ordering sampling to give itself better information about the cost, as well as the likely benefit, of the discovery sought.[162]

The second set of factors to be considered in deciding whether to sample relate to the likely content of the information sought. There are two relevant inquiries regarding this matter. First, there must be sufficient uncertainty as to the contents of the discovery in dispute, such that reducing uncertainty and obtaining better information about the likely content of the information sought will substantially aid the court in resolving the discovery dispute.[163] This factor is ultimately addressed by the subjective confidence level of the judge deciding the discovery motion.[164] In cases where a judge is confident, based on prior discovery,

158 Semsroth v. City of Wichita, 239 F.R.D. 630, 638 (D. Kan. 2006).

159 Parkdale Am., LLC v. Travelers Cas. & Sur. Co. of Am., No. 3:06CV78, 2007 WL 4165247, at *9 (W.D.N.C. Nov. 19, 2007).

160 *Zubulake I*, 217 F.R.D. at 312 n.9.

161 Note in *Zubulake*, for example, that defendants' initial estimated cost of $300,000 to restore the e-mails on the backup tapes was later reduced to $175,000. *Id.* at 312–33.

162 In *McPeek*, Magistrate Judge Facciola required that the Department of Justice, the producing party, "carefully document the time and money spent in doing the search" and "[u]pon the completion of this search . . . then file a comprehensive, sworn certification of the time and money spent and the results of the search." 202 F.R.D. 31, 35 (D.D.C. 2001).

163 *See* Brokaw v. Davol, Inc., No. 07-5058, 2011 WL 579039, at n.7 (R.I. Super. Ct. Feb. 15, 2011) (declining to sample where plaintiffs "have . . . provided the Court with examples of the information that they hope to find").

164 There is extensive literature on subjective probability judgments and confidence intervals regarding those judgments. *See* Daniel Kahneman & Amos Tversky, *Subjective Probability: A Judgment of Representativeness*, *in* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES 32 (Daniel Kahneman, Paul Slovic & Amos Tversky eds., 1982); Charles Yablon, *The Meaning of Probability Judgments: An Essay on the Use and Misuse of Behavioral Economics*, 2004 U. ILL. L. REV. 899. For our purposes, it is sufficient to note that the literature makes a distinction between an individual's probability judgment as the likelihood of the occurrence of an event and an individual's confidence that their probability judgment is correct. For example, I may believe there is a 1/6 chance that a single die will roll a six on the next roll, but my confidence in that probability judgment is close to 100%. Alternatively, I may believe the Mets have a

testimony, or the arguments of the parties, that there is likely to be relevant information in the requested files of sufficient importance to meet the proportionality test, sampling is unnecessary and inappropriate. Likewise, where the evidence (or lack thereof) has already convinced the judge that the requested discovery is unlikely to contain any information sufficient to meet the proportionality test, sampling is similarly inappropriate.[165] In short, a significant level of uncertainty as to the contents of the information sought is a precondition for sampling. That said, because we believe that sampling is a useful and flexible tool that can improve the quality of judicial decision-making, we would encourage its use whenever a judge believes that her confidence in the correctness of her decision on the motion can be significantly increased by information obtained through sampling.

This brings us to the second merits-based factor—the ability of sampling to provide information that will substantially improve the accuracy of the decision-making. Specifically, in ordering sampling, a judge must reasonably believe that the sample will substantially or wholly alleviate the uncertainty about the requested discovery. The results of the sample must give the judge markedly greater confidence that the requested discovery does or does not contain information of sufficient importance to meet the proportionality test. Put another way, there must be a reasonable likelihood that a sample of a size within the generally accepted scale[166] will produce enough information to eliminate substantial uncertainty as to the contents of the requested discovery. If such a sampling would be insufficient to resolve such questions, there is little reason to proceed with the sampling even if uncertainty is present as to the requested documents' contents.

For example, suppose the discovery dispute involves a single "smoking gun" e-mail, allegedly sent from defendant to plaintiff "sometime in the summer of 2005." Plaintiff swears he saw it (on an office computer system to which he no longer has access); defendant swears it never existed. There is undisputed testimony that the only place the e-mail might still exist is in the backup tapes of the office computer, which will cost $100,000 to fully recover and search for the

"reasonable likelihood" of being a pennant contender this year, but admit that my confidence in that probability judgment is quite low. Similarly, we believe that judges can distinguish between their assessment, at any given time, as to the likelihood of success of a discovery motion and their confidence as to the accuracy of that assessment. The factor we discuss here relates to that subjective assessment of accuracy.

165 Although, as we discuss later, the degree of unlikelihood of finding important evidence may be relevant to determining if the party seeking discovery is doing so in good faith or for harassment purposes (and therefore should be subject to cost-shifting, discovery cut-off, or sanctions), that inquiry relates primarily to the judge's determination of the subjective good faith of the party seeking discovery based on the information available to that party at the time the motion. Accordingly, obtaining new information about that discovery through sampling will usually be neither necessary nor relevant.

166 See discussion *infra* Part III.B concerning the advisable size of a sample.

period June 2005 through September 2005. While the cost criteria for sampling are arguably met, and there is substantial uncertainty regarding the merits of the underlying discovery motion, we do not think this is a strong candidate for sampling. The reason is that examination of 15% to 25% of the requested files is unlikely to sufficiently reduce the uncertainty regarding the existence of the "smoking gun" e-mail. Because it involves only a single e-mail, searching, for instance, the June 2005 e-mails and finding nothing will not substantially diminish the likelihood that the e-mail will be found in the July, August, or September files. Put another way, the three months of files that do not contain the e-mail are not expected to be representative of the files for the month in which the email may exist.[167] This example therefore provides illustration of a critical point—sampling works best, and is most likely to be ordered, when there is good reason to believe that the contents of the materials sampled are representative or over-representative[168] of the contents of the entire set of materials sought in discovery. It is only when such representativeness exists that sampling is likely to be a cost-effective way to provide increased accuracy in decision-making. The importance of this factor reflects our view that sampling serves an information-gathering function and should not be thought of as a mere incrementalization of discovery. If it cannot provide substantially improved information to a court in a cost-effective manner, it is of little utility.[169]

In sum, courts should order sampling when there is a significant level of uncertainty as to the contents of the information sought, the size of the litigation and cost of discovery are sufficiently large, and when sampling has the potential to substantially alleviate the uncertainty surrounding the requested discovery. As to the procedures for ordering

---

167 Of course, from a purely mathematical point of view, the fact that the e-mail did not exist in the June files makes it somewhat less likely that it exists at all, but that result is also consistent with a belief to a high degree of probability that it exists in the files of one of the other months. In this context, Bayes Theorem may be of use in deriving conditional probabilities—that is, probabilities that a given fact about the world is true (e.g., "those files contain important new information") given the observation of some other fact or piece of evidence (e.g., "the sample does not contain any important new information"). Proponents of the use of Bayes Theorem argue that it provides a rational method for arriving at probability judgments with respect to both objective, frequentist probabilities, and subjective probability judgments like those judges make in deciding individual cases. There are ongoing arguments among scholars as to whether Bayes Theorem accurately depicts decision-making and should be used more frequently in deciding legal issues. For the purposes of this Article, we are agnostic on all these issues. Nonetheless, we believe that Bayes Theorem can be useful in illustrating certain points made earlier in this paper by providing mathematical models of different examples of decision-making by judges utilizing the sampling procedure. The equations and examples of such an application are set out in Appendix A, *infra*.

168 The significance of over-representation is discussed *infra* Part III.B.

169 A court conceptualizing sampling as a mechanism to phase discovery in order to constrain costs—an approach we do not endorse—would not likely consider this factor, as the phased approach does not concentrate on overcoming information deficits but rather is focused exclusively on cost minimization. *See infra* Part I.C.

sampling, we believe, as previously stated, that sampling is a technique that judges may or may not choose to utilize. As such, its use is solely within the judge's discretion and may never be claimed by a party as of right. Nonetheless, parties either making or resisting discovery motions are free to suggest to the court that sampling might be an appropriate technique for resolving difficult discovery issues, and might even provide data concerning costs of discovery, costs of sampling, suggested sampling techniques, etc.

Such suggestions for sampling will frequently be a tactical two-edged sword. As we have seen, suggesting sampling means that you believe the court is experiencing a significant level of uncertainty as to whether to grant or deny the motion (something litigants are frequently reluctant to concede), and that information obtained from sampling will help resolve the dispute. On the other hand, when both parties feel they have a great deal riding on a discovery motion, the outcome of which is uncertain, they may find it in their mutual interest to jointly suggest not only that the court order sampling prior to deciding the motion, but also might be able to agree on a proposed sampling protocol that balances the cost-saving and information-gathering aspects of the sampling technique to the satisfaction of both parties. Such agreements as to sampling would not only represent useful examples of cooperation among parties with respect to electronic discovery, but are also very likely to be accepted by the court.

Whether the parties agree or disagree as to the utility of sampling in their initial motion papers, they should each be free to make additional arguments regarding that motion once the results of such sampling are known. This would be in accordance with the two-part briefing procedure we suggest in the following Sections.

B. *Method of Sampling*

Our theoretical discussion in Part I argued that since sampling is best understood as a cost-effective technique for more accurately resolving discovery disputes, the optimal sample size is the one that provides the most additional information to the court at the lowest cost. Our empirical analysis in Part II has revealed that in cases in which the cost of sampling exceeded 25% of the total cost of discovery, courts did not order sampling. Likewise, where the costs of sampling would be less than or equal to 25% of the total cost of discovery, a sample was ordered. Among those cases in which sampling was ordered, the sample on average was 14.46% of the total cost of the discovery. Thus, it would appear that courts have adopted a 25% threshold with a preference for ordering a sample consisting of about 15% of the total cost of

discovery.[170]

While this 15% to 25% sample size can be viewed as another prerequisite for the ordering of sampling, the size of the sample to be ordered is obviously malleable in a way that other sampling criteria are not. That is, while the judge, in making her sampling order, cannot affect the amount in controversy in the litigation, the cost of discovery sought, or even the degree of uncertainty that exists with respect to the importance of the information sought, she can, and indeed must, determine what percentage of the information will be subject to sampling. In setting the parameters and size of a sample, a court should adhere to the general principle of trying to maximize information at the lowest possible cost.

This makes especially salient the facts found in Part II, that in cases in which sampling is ordered, the average sample costs 15% of the total cost of discovery. Further, we also found that 25% is the upper limit for ordering sampling, with no court ordering sampling at a cost exceeding that threshold. This data strongly suggests that the courts themselves feel that the sampling technique ceases to be cost effective at percentages above 25%, and that a range of about 15% of the costs of discovery sought is likely to be optimal. These percentages certainly make sense in light of the criteria we have previously outlined for use of the sampling technique. Samples much below 15% would be unlikely to provide confidence that they accurately reflect all the types of information contained in the files sought.[171] Samples above 25% are likely to involve added costs for diminishing returns. Accordingly, while we believe there should be no absolute minimum or maximum cost threshold for the size of a sample, we view the 15% to 25% range as a useful guideline for ensuring the sample provides adequate information about the totality of the requested documents while not imposing unnecessary costs.

Note also that our empirical study and our proposed guidelines both look at the sample size in terms of relative cost of discovery rather than quantity of material reviewed. Of course, courts that have ordered sampling have not always conceived of sample size in cost-based terms. Courts will often order sampling based on a percentage of the total number of backup tapes at issue. Yet we believe thinking about the size of sampling in monetary terms makes good sense, particularly with respect to electronic discovery. This is because the marginal cost of producing backup tapes decreases with each increase in the number of

---

170 *See* AAB Joint Venture v. United States, 75 Fed. Cl. 432, 444 (Fed. Cl. 2007) ("The Court believes that restoration of one-fourth of the total back-up tapes should be adequate to determine whether the tapes are likely to possess relevant evidence.").

171 It is true that randomized scientific surveys can often provide accurate data regarding certain characteristics of a population based on much smaller sample sizes, but, as we discuss in the next Section, such scientific surveys usually have a more limited purpose, and one that is not exactly the same as a purpose of most sampling inquiries.

tapes because there are fixed costs relating to production that are imposed irrespective of the number of tapes that are ultimately produced. In that sense, the first tape is always the most costly. Accordingly, basing the size of the sample on a percentage of the total number of backup tapes instead of the total cost of discovery may yield sampling that is not fully representative of the costs related to discovery. Some courts may have shied away from basing size in monetary terms because courts lacked sufficient information to issue an order based on such a metric. Our suggested approach would have the court seek to ascertain, prior to issuing a sampling order, whether information on marginal per unit costs is available and if so, to utilize that in the sampling order. If that information is not readily available, the court should ascertain whether it can be developed as part of the sampling inquiry itself. The fact that most sampling decisions thus far have ordered samples in the 15% to 25% range based on discovery costs, strongly suggests cost is the appropriate criteria for evaluating sample size.

In Part II, we also noted that courts have utilized a number of different and inconsistent methods for determining how the sample files or information are selected for review. We divided the cases into four broad categories of methodology:

1) *Best Case Scenario*, in which the party requesting documents selects the backup tapes or files it believes are most likely to contain important information that would satisfy the proportionality standard. The requesting party has exclusive control over choosing which files are part of the sample, subject only to sample size limitations imposed by a court.

2) *Scientific Sampling*, in which the court orders a sample based on randomized selection of files or documents so as to produce a statistically reliable sample whose characteristics are likely to reflect that of the larger population from which it was selected at some statistical confidence interval.

3) *Parties Stipulation*, in which the sample reflects an agreement between the parties as to the amount and type of information or files that will be sampled. This approach is functionally equivalent to private agreements to sample, subject to limited oversight by courts.

4) *Court Order*, in which the court simply orders a sample of a particular size and type, and which may reflect all, some, or none of the considerations involved in Best Case Scenario, Scientific Sampling, and Parties Stipulation methodologies.

Our empirical study found that recent cases have used all four methodologies, and that the choice of methodology did not appear to be correlated with any other independent variable to a statistically significant degree. This was also consistent with the language of the cases themselves, which rarely discussed the question of which

methodology to use or sought to defend the particular methodology selected. Nonetheless, our empirical study did show that far more judges opted for the "best case scenario" or "court order" methodology[172] than sought to utilize "scientific sampling."[173] This may appear surprising given the vast importance of statistical analysis in most areas of academic social scientific inquiry, increasingly including law.[174] Moreover, some of the cases utilizing scientific sampling were among the few that attempted to justify their choice of methodology.[175]

While the apparent popularity of the best case scenario and court order methodologies may reflect, in part, the enormous influence of the *Zubulake* case (which can be seen as a application of both those methods), we believe their popularity, and the reluctance of courts to employ scientific sampling methods, is not only understandable, but, in many cases, normatively preferable. Often it will be "best case scenario" not "scientific" sampling that is likely to provide the most useful information to the court at the lowest cost. The reason for this is simple. Scientific sampling is a statistical methodology whose fundamental goal is to ensure that the sample analyzed is representative of the larger population about which information is sought. In the context of resolving a discovery dispute, however, a judge may rationally seek a sample that is over-representative of the total population with respect to documents with particular characteristics. That is, if the judge is seeking to determine whether the information contained in the discovery sought is likely to contain documents of sufficient importance to justify the costs of discovery, a sample with a greater than average number of the most important documents may be more helpful in making that determination than a scientifically selected, representative sample. This is because the over-representative sample provides the court with more information about the particular documents about which it is most interested, the ones most likely to be important to the case. A scientific sample, in contrast, would provide fewer of the important documents, but at the same time provide a more reliable estimate of the total number of such documents that are likely to be in the discovery sought.

A simple example can illustrate the trade-off involved. Assume you are a magazine writer, thinking about researching and writing a story

---

172 Ten of the cases in the study (31.25%) utilize "best case scenario" methodology. Another twelve (37.5%) involved a "court order" methodology.

173 Only three cases (9.38%) utilized "scientific sampling."

174 *See generally* DAVID W. BARNES & JOHN M. CONLEY, STATISTICAL EVIDENCE IN LITIGATION: METHODOLOGY, PROCEDURE AND PRACTICE (1986); MORRIS DEGROOT, STEPHEN E. FIENBERG & JOSEPH B. KADANE, STATISTICS AND THE LAW (1986); STEPHEN E. FIENBERG, THE EVOLVING ROLE OF STATISTICAL ASSESSMENTS AS EVIDENCE IN THE COURTS, REPORT OF THE PANEL ON STATISTICAL ASSESSMENTS AS EVIDENCE IN THE COURTS (1989); MICHAEL O. FINKELSTEIN, QUANTITATIVE METHODS IN LAW: STUDIES IN THE APPLICATION OF MATHEMATICAL PROBABILITY AND STATISTICS TO LEGAL PROBLEMS (1978).

175 *See, e.g.*, *In re* Vioxx Prods. Liab. Litig., Nos. 06-30378, 06-30379, 2006 WL 1726675, at *2 n.5 (5th Cir. May 26, 2006) (per curiam).

about octogenarian surfers in California for *Sports Illustrated*. You think the story could be worth investigating, but you are unsure that there are any octogenarian surfers, and if so, whether they are sufficiently interesting to make a good story. One way to research the story (assuming you are stuck on the east coast) would be to get a list of California octogenarians (perhaps from the California AARP), call a substantial number at random, and ask whether any of them surf. Assuming the sample were large enough, this would enable you to estimate, with scientific accuracy, the number of octogenarian surfers in the state of California. Unfortunately, it would probably also result in your speaking to very few octogenarian surfers. A better approach might be to limit your sample to octogenarians with addresses in Malibu, on the theory that most octogenarian surfers, and certainly most interesting octogenarian surfers, are likely to live in Malibu and similar communities. The over-representativeness of the sample is precisely the point. You will be able to find out more information about more octogenarian surfers for the same expenditure of time and effort. This will enable you to make a better determination as to whether the story is worth pursuing. One thing you will not be able to do, however, based on your skewed sample, is accurately estimate the total number of octogenarian surfers in California, but that was not the purpose of your inquiry.

Judges considering use of sampling are also generally looking to gather as much information as possible about documents with unique characteristics, i.e. importance to the case, which like octogenarian surfers, are at best rare within the population being searched, and may not exist at all. In such circumstances, if there is good reason to believe that one segment of the population for which discovery is sought is likely to contain an over-representative sample of such documents, the principle of maximizing information to the judge in the most cost-effective manner would require that the over-representative, "best case scenario" approach be applied.

This analysis not only explains our finding of a preference for "best case scenario" over scientific methodology in most of the reported cases, it also explains the use of the scientific approach in a minority of cases. In the three cases in our survey in which scientific methods of sampling were utilized or endorsed, the purpose for which the sample was sought was quite different from most of the other cases. It was to test the accuracy or reliability of certain word searching protocols to ensure that the populations being sampled actually contained documents with the characteristics they were supposed to contain. For example, *D'Onofrio v. SFX Sports Group, Inc.*[176] involved a dispute over a privilege log that listed 9,413 documents withheld for privilege, work product, or other

176 256 F.R.D. 277, 278 (D.D.C. 2009).

reasons. At a hearing, it was agreed that plaintiff would be allowed to "test the validity of the privilege log using statistical sampling." Plaintiff's expert would be permitted to "select a representative sample, that would be made available to plaintiff's counsel for his review to determine whether the privileges asserted were in fact appropriate."[177] Note that in *D'Onofrio*, unlike *Zubulake* and similar cases, the judge was not interested in finding out as much as possible about a small subset of important documents within the population of documents sought to be discovered. Rather, he was concerned with whether defendant's characterization of all the documents in the privilege log as privileged was accurate, and estimating, as accurately as possible, how many errors the privilege log contained. As we have noted, making such a determination requires scientific, statistical methods, precisely what the court in *D'Onofrio* ordered. The other cases utilizing or endorsing scientific sampling methodology involve similar concerns. They seek to use scientific sampling methods to test whether given word search protocols reliably produced populations of documents with the characteristics the parties claimed.[178]

A number of general principles emerge from the foregoing analysis:

1) In situations such as *Zubulake*, where the court is seeking as much information as possible about the likely contents of the documents being sought in order to determine their importance to the litigation, the court should seek to select a sample which is most likely to contain the largest number of such important documents.

2) This is preferably done by agreement of the parties, recognizing that the party seeking discovery generally has an interest in producing a sample that has the greatest likelihood of containing important documents and the party opposing discovery generally has an interest in minimizing the costs of the sample. Nonetheless, if the parties fail to reach agreement, or if the judge believes the sample recommended by the parties will not provide maximum information about the contents of the requested discovery at minimum cost, the judge may order a different sample, based either on recommendations of the party seeking discovery alone, or on the basis of any other information available to the court.

---

177 *Id.* at 279.

178 *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 256–57 (D. Md. 2008) (criticizing the defendant's keyword search for privileged documents for not being tested by statistical sampling). "The only prudent way to test the reliability of the keyword search is to perform some appropriate sampling of the documents determined to be privileged and those determined not to be in order to arrive at a comfort level that the categories are neither over-inclusive nor under-inclusive." *Id.* at 257; *see also* William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 136 (S.D.N.Y. 2009) (stating that "common sense" requires that sampling must be employed in determining the completeness of keyword search protocols).

3) However, if the purpose of the sample is not to obtain as much information as possible about the likely contents of the documents being sought in order to determine their importance to the litigation, but rather, to determine the extent to which an entire set of documents being sought or withheld from discovery actually have a certain characteristic (e.g. privilege, relevance) then scientific sampling techniques based on representative samples should be utilized.

A final methodological issue—one that has received almost no discussion from courts[179]—is which party should pay for the sampling. For a number of reasons, we believe that the correct approach is for the responding party (the party opposing discovery) to pay for the actual production of the sample. First, it appears from our survey that this is the procedure currently being followed by courts that order sampling. Second, as we noted in Part I, the basic presumption of the federal discovery rules remains that the producing party bears its own costs.

Although we have resisted the notion that sampling represents partial or incremental discovery, our analysis assumes that the judge is considering whether this is a case in which the traditional presumption should be altered (that is why she is utilizing sampling) but *has not yet made any such determination*. Accordingly, to shift costs at the sampling stage would be premature. However, there may be unusual circumstances in which the results of the sampling or other facts might lead courts to shift the costs of sampling to the requesting party.[180]

C. *Briefs Regarding Sampling*

As previously noted, the original suggestion for sampling may come either from the court or from one or both of the parties. In either event, however, it is likely that the parties will have important information that will aid the court in determining whether to sample and shaping the sampling methodology. Accordingly, we envision a two

179 *See supra* notes 128, 130.

180 FED. R. CIV. P. 37(a)(5) does, of course, contain provisions that presumptively shift the cost of making discovery motions, including attorneys' fees, on the party losing such motions. Such cost-shifting does not apply when the losing party's position is "substantially justified" or when cost shifting is otherwise unfair. In situations where sampling is ordered, as we have seen, there must be sufficient uncertainty as to the appropriate result, such that both parties' positions are likely to be substantially justified. Nonetheless, in a case in which sampling or other evidence reveals to the court that the party seeking discovery was essentially engaged in an abusive fishing expedition having no substantial justification for the discovery sought, then the cost-shifting provisions of FED. R. CIV. P. 37(a)(5) should result in a shifting of the cost of the sample as well as the other costs of the discovery motion. In making such a determination, as in other areas of the law, courts must be wary of the danger of "hindsight bias." *See generally* Baruch Fischhoff, *For Those Condemned to Study the Past: Heuristics and Biases in Hindsight*, *in* JUDGMENT UNDER UNCERTAINTY: HEURISTICS AND BIASES, *supra* note 164, at 335; Jeffrey J. Rachlinski, *A Positive Psychological Theory of Judging in Hindsight*, 65 U. CHI. L. REV. 571, 608–15 (1998).

stage briefing procedure. The first stage of briefing would occur prior to the ordering of sampling and focus on whether to sample and how to shape the sampling protocol. Because, as previously noted, the parties may be reluctant initially to endorse sampling, and because sampling is ultimately the judge's own "discovery about discovery," the judge should feel free to direct the parties briefing to the particular issues on which she wants more information. For example, in considering application of a "best case scenario" methodology, the court might direct the moving party to submit a brief not just on whether sampling should be ordered, but, *if* the court were to order sampling, describing what that party would consider an optimum sample and why they believe the results of that sample will support their position with respect to the underlying discovery motion. The parties could be asked to confer on these proposals in an attempt to reach agreement on a sampling protocol, and if they cannot, the responding party would reply, perhaps with a counterproposal of its own. The responding party would also be expected to supply additional information regarding costs of production of requested samples of different sizes. Such information is often provided by expert affidavits or testimony. The other side could then respond to those cost estimates. At the end of this process, there might be an agreement between the parties or one or more sampling proposals tied to the costs and needs of the individual case. The court would then choose one of these sampling protocols or create its own based on the principle of maximizing information to the court in the most cost-effective manner.

The second phase of the briefing would come after sampling has occurred and the parties have had a chance to review the sampled documents. These briefs would seek to show how the results of the sampling support the parties' respective positions on the underlying discovery motion.

D. *Interpreting and Applying the Results of Sampling*

The final unanswered question revealed by our study is when the results of sampling can support a decision to shift costs and when it can support an order cutting off further discovery. Both results are authorized under the Federal Rules in situations where the proportionality test is not met, and both results have been ordered by courts after analyzing the results of sampling. Indeed, of the two paradigmatic sampling cases, *McPeek* and *Zubulake*, the latter involved cost-shifting while the former resulted in a discovery cut-off. Neither case, nor the others in our sample, provides much explicit guidance on when each of these orders is appropriate. It appears that whether discovery costs were shifted or discovery cut-off frequently depended on

which result was requested by the party opposing the discovery.

We believe that the greater, more nuanced information made available by sampling makes it unnecessary for the courts to choose between these two potential orders until it has analyzed the results of the sample. Specifically, we believe that whenever a court must decide whether a given discovery request meets the proportionality test, it should consider three possible results. When sampling reveals a substantial likelihood of important information in the requested discovery, making it likely that the requested discovery satisfies the proportionality standard, the court should order the discovery with production costs on the responding party. When sampling reveals little or no substantial likelihood of important information in the discovery sought, the court must make a finer-grained analysis. In circumstances where there is still some possibility the requested discovery might contain useful information and where the burden on the responding party can be largely or completely alleviated by having the requesting party bear the costs of production, cost-shifting should be ordered. This is a recognition that cost-shifting is a less preclusive and therefore generally preferable response to a discovery cut-off. Cost-shifting is less preclusive because it can be applied to various degrees (as the court did in *Zubulake*) and because it interferes less with the losing parties' ultimate chance of success on the merits. A less preclusive remedy is preferable because, as we discussed in Part I, the court is still making merits-based determinations about the case on less than a full record and such prejudging should always be as tentative and reversible as possible. Cost-shifting, either in whole or in part, does not foreclose the discovery sought, but it does force the losing party to consider just how valuable that discovery is, and how much he wants to invest in furthering his cause. Moreover, the responding party, whose costs are now substantially covered, has far weaker grounds to object to producing information he claims has no importance to the case.

Accordingly, while cost-shifting should be the applicable result for most requested discovery where sampling indicates the proportionality standard has not been met, the more preclusive option to cut-off discovery completely is also available to the court. Such discovery cut-offs should be considered when the sample demonstrates that there is no reasonable likelihood of finding relevant information in the discovery sought, and that cost-shifting alone will not fully alleviate the burden on plaintiff of making production. While not quite a finding of discovery abuse or "fishing expedition,"[181] this would amount to a finding that further discovery would serve no useful purpose. Although, given our

[181] We believe that it is only upon a finding that a party has engaged in such abusive discovery that a court should consider an ex post reversal of the costs of the sample production. *See supra* note 180.

general concerns about prejudgments on the merits, such orders should be made sparingly, we also believe the added information available to the court through sampling will enable it to make such orders with assurance in the appropriate case.

CONCLUSION

As the first extensive theoretical and empirical analysis of sampling in e-discovery disputes, we intend this Article to be the beginning of a discussion among lawyers, judges, and academics. Accordingly, rather than conclude with a summary of previously made points, we prefer to conclude with the following two appendices, which we hope will promote further discussion of sampling as an important and rapidly developing feature of e-discovery. The first appendix illustrates some of our major points regarding the utility of sampling through the application of Bayes Theorem, showing that our arguments are consistent with Bayesian analysis, although not dependent on them. Our second appendix is a preliminary attempt at a recommended set of "best practices" for sampling procedures, which we hope will form the basis for further debate and refinement.

APPENDIX A: AN APPLICATION OF BAYES THEOREM TO SAMPLING DECISIONS

Bayes Theorem may be of some usefulness in analyzing conditional probabilities in a judge's decision-making process.[182] What follows are mathematical models of different examples of decision-making by judges utilizing the sampling procedure. The analysis of these examples are based on the following equation:

$$\Pr(A \mid X) = \frac{\Pr(X \mid A)\Pr(A)}{\Pr(X \mid A)\Pr(A) + \Pr(X \mid \sim A)\Pr(\sim A)}$$

For our purposes in these examples, the variables can be interpreted as follows:

Pr(A|X) = Likelihood that no important documents exist in the files, given that no important documents were found in the sample. This is the determination the judge ultimately seeks to make.

Pr(X|A) = Chance of a negative finding in the sample given that no important documents exist in the files. This will always be true in sampling cases, thereby giving this a value of 1. Accordingly, in our Bayesian analysis, the numerator will always equal Pr(A)

182 *See supra* note 167.

Pr(A) = Likelihood that no important documents exist in the files sought to be discovered.

Pr(~A) = Likelihood that important files do exist in the files sought to be discovered

Pr(X|~A) = Likelihood of a negative sampling result even if important documents exist in the files sought to be discovered. We can call this a "false negative," the only kind of false result that can exist in this discovery context.

Using this equation we evaluate the usefulness of sampling in a number of different situations:

Example A: Little Uncertainty As to Motion, Sample Marginally Helpful in Reducing Uncertainty

Consider the situation discussed *infra* in Part III.C where the judge strongly believes there is one, and only one, critical e-mail in the files, and it exists either in the June, July, August, or September files of the defendant. We said that it made little sense to sample in this case, since a negative finding with respect to June would do little to affect the judge's view of the existence of important information. We can model the strong belief that a critical e-mail exists in the files by giving Pr(A) and Pr(~A) values of 0.1 and 0.9 respectively. If we assume the judge believes that critical email is equally likely to be found in the e-mails of any of the summer months, then the value of Pr(X|~A), the likelihood that the sample will show up negative even if the critical e-mail exists, is 0.75. Plugging these numbers into the equation gives us a value for Pr(A|X), the "posterior probability," of 12.9%. In other words, the judge's belief that there might not be important evidence in the entirety of the files sought has gone from 10% to slightly under 13%. Clearly, such sampling will not have a significant effect on the outcome of the discovery motion.

Example B: Uncertainty As to Motion, Sample not Significantly Helpful in Reducing Uncertainty

To model a judge who is uncertain regarding the discovery motion, but is leaning slightly toward granting it, we assigned prior probabilities Pr(A) and Pr(~A) of 0.4 and 0.6 respectively. To model a sample that is not strongly helpful in reducing uncertainty, we assumed a 50% likelihood of a false negative. These numbers give us a posterior probability of 57%. The judge's belief that there might not be sufficiently important evidence in the entire file to justify granting the motion has changed from weakly negative (40%) to weakly positive (57%). A judge who knew beforehand that the results would be this inconclusive might

not bother to order sampling at all. On the other hand, judges frequently do not know what all the possible results of sampling might be, and therefore cannot know the true value of Pr(X|~A)[183] prior to ordering the sample. Obtaining such a result, and therefore now believing that the evidence weakly favors the party objecting to discovery, the court might incline toward a weakly preclusive result, like shifting the costs of some percentage of the requested discovery.

Example C: Uncertainty As to Motion, Sample Significantly Helpful in Reducing Uncertainty

Here again, the judge's uncertainty is modeled as prior probabilities Pr(A) and Pr(~A) of 0.4 and 0.6 respectively (slightly leaning towards granting the discovery). However, to model a sample that is of significant help in reducing uncertainty, we assume only a 20% likelihood of a false negative. That is, if the files contain the information the party seeking discovery claims is there, some of that information is very likely to show up in the sample. In these circumstances, failing to observe any such data in the sample will yield a posterior probability of 76.9% that the requested files do not contain such data (even though the judge was mildly positive toward the motion to begin with). Here, sampling clearly serves its purpose of giving the judge greater insight and certainty concerning the motion pending before her.

APPENDIX B: PROPOSED "BEST PRACTICES" FOR THE USE OF SAMPLING IN THE RESOLUTION OF DISCOVERY DISPUTES

I. THE DECISION TO ORDER SAMPLING

Courts should order sampling when 1) the size of the litigation and cost of discovery are sufficiently large, 2) a significant level of uncertainty exists as to the contents of the information sought, and 3) sampling has the potential to substantially alleviate the uncertainty surrounding the requested discovery.

There is no strict cut-off in the size of the litigation or in the absolute or relative cost of discovery necessary to justify sampling.

---

183 In this simplified model, there is a single sampling result *X*, which represents a finding of no important evidence. In the real world, of course, real sampling can produce many different potential results, such as the *Zubulake* finding of some marginally relevant evidence, but no "smoking guns." Because the degree to which a sampling result will alleviate uncertainty frequently cannot be known in advance, sampling such as the one in this model may frequently be ordered. Once the sample has been taken, of course, the range of possible outcomes becomes narrowed to one, and it is possible to develop a posterior probability based on the actual outcome of the sample as indicated above.

However, discovery costs of $20,000 or less have been held not to warrant sampling, while discovery estimated at $175,000, which was 1.35% of plaintiff's maximum compensatory recovery of $13 million, was considered an appropriate candidate for the sampling technique.

In cases where the actual costs of the discovery sought are substantially disputed or otherwise uncertain, sampling may provide a more reliable cost estimate.

There must be sufficient uncertainty as to the contents of the discovery in dispute that reducing that uncertainty and obtaining better information about the likely content of the information sought will substantially aid the court in resolving the discovery dispute. This factor is ultimately addressed to the subjective confidence level of the judge deciding the discovery motion.

A judge must reasonably believe that the sample will substantially or wholly alleviate the uncertainty about the requested discovery. Sampling should be used whenever a judge believes that her confidence in the correctness of her decision on the motion can be significantly increased by information obtained through sampling.

II. PROCEDURES FOR ORDERING SAMPLING

The decision to order sampling is solely within the judge's discretion and may never be claimed by a party as of right. Nonetheless, parties either making or resisting discovery motions are free to suggest to the court that sampling might be an appropriate technique for resolving difficult discovery issues, and may provide data concerning costs of discovery, costs of sampling, suggested sampling techniques, etc.

Ordinarily, resolution of discovery disputes through sampling will involve a two-stage briefing procedure. The first stage of briefing would occur prior to the ordering of sampling and focus on whether to sample and how to shape the sampling protocol. The judge should feel free to direct the parties briefing to the particular issues on which she wants more information.

The second phase of the briefing would come after sampling has occurred and the parties have had a chance to review the sampled documents. These briefs would seek to show how the results of the sampling support the parties' respective positions on the underlying discovery motion.

III. METHODS OF SAMPLING

Courts should utilize a sample size that maximizes information to the judge at the lowest possible cost. While there is no absolute

minimum or maximum cost threshold for the size of a sample, prior judicial practice indicates that a sample of 15% to 25% of the total cost of the discovery sought on the underlying motion provides adequate information about the totality of the requested documents while not imposing unnecessary costs.

If possible, sample size should be based on a percentage of the total costs of the discovery sought. The court should seek to ascertain, prior to issuing a sampling order, whether information on marginal per unit costs is available and if so, to utilize that in the sampling order. If that information is not readily available, the court should ascertain whether it can be developed as part of the sampling inquiry itself.

In situations where the court is seeking as much information as possible about the likely contents of the documents being sought in order to determine their importance to the litigation, the court should seek to select a sample which is most likely to contain the largest number of such important documents. This is preferably done by agreement of the parties, recognizing that the party seeking discovery generally has an interest in producing a sample that has the greatest likelihood of containing important documents and the party opposing discovery generally has an interest in minimizing the costs of the sample. Nonetheless, if the parties fail to reach agreement, or if the judge believes the sample recommended by the parties will not provide maximum information about the contents of the requested discovery at minimum cost, the judge may order a different sample, based either on recommendations of the party seeking discovery alone, or on the basis of any other information available to the court.

If the purpose of the sample is not to obtain as much information as possible about the likely contents of the documents being sought in order to determine their importance to the litigation, but rather to determine the extent to which an entire set of documents being sought or withheld from discovery actually have a certain characteristic (e.g., privilege, relevance, etc.) then scientific sampling techniques based on representative samples should be utilized.

The producing party should pay the costs of producing the sample, unless the sample results or other subsequent information reveals that the underlying discovery motion is subject to the cost-shifting provision of FED. R. CIV. P. 37(a)(5).

IV. APPLYING THE RESULTS OF SAMPLING

When sampling reveals a substantial likelihood of important information in the requested discovery, the court should order the discovery with production costs on the responding party.

When sampling reveals little or no substantial likelihood of

important information in the discovery sought, but there is still some possibility that the requested discovery might contain useful information, and where the burden on the responding party can be largely or completely alleviated by having the requesting party bear the costs of production, cost-shifting should be ordered.

When sampling demonstrates that there is no reasonable likelihood of finding relevant information in the discovery sought, and that cost-shifting alone will not fully alleviate the burden on plaintiff of making production, a discovery cut-off may be ordered.