Vanderbilt University Law School

Scholarship@Vanderbilt Law

Vanderbilt Law School Faculty Publications

Faculty Scholarship

2012

When 10 Trials are Better than 1000: An Evidentiary Perspective on Trial Sampling

Edward K. Cheng

Recommended Citation

University of Pennsylvania
Law Review

ESSAY

WHEN 10 TRIALS ARE BETTER THAN 1000: AN EVIDENTIARY PERSPECTIVE ON TRIAL SAMPLING

EDWARD K. CHENG[†]

INTRODUCTION

In many mass tort cases, individual trials are simply impractical. Take, for example, *Wal-Mart Stores, Inc. v. Dukes*,[1] a class action em-

ployment discrimination suit that the Supreme Court reviewed last Term. With over 1.5 million women potentially involved in the litigation,[2] the notion of holding individual trials is fanciful. Other recent examples of the phenomenon include the *In re World Trade Center Disaster Site Litigation*[3] and the fraud litigation against light cigarette manufacturers,[4] in which Judge Weinstein colorfully noted that any "individualized process . . . would have to continue beyond all lives in being."[5]

Faced with an unserviceable number of plaintiffs, courts have proposed sampling trials: rather than litigating every case, courts would litigate a small subset and award the remaining plaintiffs statistically determined amounts based on the results. But while sampling is standard statistical practice and often accepted as evidence in other legal contexts,[6] appellate courts have balked—based on due process concerns—at the notion of court-mandated, binding trial sampling.[7]

---

[1] 131 S. Ct. 2541 (2011).

[2] *Id.* at 2547. The Supreme Court rejected class certification on commonality grounds, *see id.* at 2554-57, and thousands—if not millions—of individual claims remain.

[3] *See In re* World Trade Ctr. Disaster Site Litig., 598 F. Supp. 2d 498, 503-05 (S.D.N.Y. 2009) (ordering a complex procedure to select a representative sample of cases for full discovery and early trial from the more than 9000 total cases filed).

[4] Schwab v. Philip Morris USA, Inc., 449 F. Supp. 2d 992 (E.D.N.Y. 2006), *rev'd sub nom.* McLaughlin v. Am. Tobacco Co., 522 F.3d 215 (2d Cir. 2008).

[5] *Id.* at 1247.

[6] *See, e.g.,* Shari Seidman Diamond, *Survey Research* (discussing the use of sampling surveys in a variety of contexts, including employment discrimination, juror bias, obscenity cases, and trademark litigation), *in* 1 MODERN SCIENTIFIC EVIDENCE: THE LAW AND SCIENCE OF EXPERT TESTIMONY 479, 483-86 (David L. Faigman et al. eds., 2010–2011 ed. 2010). *See generally* 2 JOSEPH L. GASTWIRTH, STATISTICAL REASONING IN LAW AND PUBLIC POLICY 483-528 (1988) (describing other uses of sampling in the legal system, including the adjudication of false labeling and gender-based equal protection claims).

[7] *See, e.g., McLaughlin,* 522 F.3d at 220, 231 (rejecting the district court's proposal to use sample trials to determine aggregate liability in litigation over whether "light" cigarettes were advertised deceptively); Cimino v. Raymark Indus., 151 F.3d 297, 319-20 (5th Cir. 1998) (rejecting the use of sampling in asbestos cases). *But see* Dukes v. Wal-Mart Stores, Inc., 603 F.3d 571, 625 n.53 (9th Cir. 2010) (noting an exception to the preference for individual hearings as a means to determine each claimant's damages when individualized evidence is difficult to obtain), *rev'd,* 131 S. Ct. 2541 (2011); Hilao v. Estate of Marcos, 103 F.3d 767, 786 (9th Cir. 1996) (noting that the sampling "methodology in determining valid claims [was] unorthodox . . . [but could] be justified by the extraordinarily unusual" circumstances). *See generally* Laurens Walker & John Monahan, *Sampling Evidence at the Crossroads,* 80 S. CAL. L. REV. 969, 974-79 (2007) (providing an overview of mass tort cases that used sampling and arguing that Judge Weinstein's proposed use of sampling in *McLaughlin* marked a step toward more efficient damage awards).

Despite this appellate reluctance, the controversy continues una-
bated. Trial courts have soldiered on by using nonbinding sampled
trials (dubbed "bellwether trials") to induce settlement,[8] and a few
brave appellate courts, including the Ninth Circuit in *Dukes*, have even
hinted at an increased receptivity to sampling.[9] Given that trial courts
have few practical alternatives, one wonders if it is just a matter of time
before their appellate brethren recognize the necessity of sampling.

The most common—and most salient—argument supporting trial
sampling is economic efficiency. Since the legal system lacks the re-
sources to litigate hundreds of thousands of asbestos cases, some kind
of resolution seems better than none. Otherwise, the tort system's
primary goals of deterrence and compensation will be profoundly un-
dercut. Opponents' objections predictably take a liberty- or rights-based
approach: defendants are entitled to individual trials, and approximate
justice will not do, no matter what the social costs.

Since sampling is nominally a species of civil procedure, the focus
on efficiency and individual rights is understandable. In this Essay,
however, I want to explore the sampling controversy from an eviden-
tiary perspective.[10] Putting aside economic and liberty interests, what
effect does sampling have on *accuracy?* Most discussions on this topic
implicitly assume that sampling is a "second best" solution, contem-
plated only in exceptional circumstances. Individual trials are the
foundation of the legal system, and though imperfect, they are pre-
sumed to be the best we can do. Thus, if we could actually try all 1.5
million cases in *Dukes*, we should. After all, since sampling involves
estimating liability from a selected subset of cases, it would appear
suboptimal to individualized adjudication. Or is it?

In the pages that follow, I offer three ways in which this "second
best" assumption can be wrong. Given the right conditions, sampling
can actually produce more accurate outcomes than individualized
adjudication. Intuitively, sampling's advantages come from its ability
to borrow strength from the different cases in the sample. Individual
adjudication confines itself to a single case and factfinder; sampling

---

[8] *See* Alexandra D. Lahav, *Bellwether Trials*, 76 GEO. WASH. L. REV. 576, 581 (2008).

[9] *See Dukes*, 603 F.3d at 625, 627 (expressing "no opinion" about the district court's
specific sampling plan, but referring to sampling as "one method" to manage damage
calculations in a large class action).

[10] For two earlier cross-cutting discussions of these issues, see Robert G. Bone, *Sta-
tistical Adjudication: Rights, Justice, and Utility in a World of Process Scarcity*, 46 VAND. L.
REV. 561, 598-617 (1993), and Michael J. Saks & Peter David Blanck, *Justice Improved:
The Unrecognized Benefits of Aggregation and Sampling in the Trial of Mass Torts*, 44 STAN. L.
REV. 815, 833-41 (1992).

does not. This basic principle recurs in each Part of this Essay, which address averaging, shrinkage, and nonrandom sampling. What's more, the good news is that the sampling procedures proposed by courts frequently capture these advantages, even if the original impetus may have been cost reduction and not accuracy.[11]

I. AVERAGING

As noted in the Introduction, the primary motivation for sampling cases is to reduce litigation costs. Sampling is necessary in asbestos, the World Trade Center litigation, light cigarettes, or any other major mass tort because courts are not equipped to run hundreds of thousands of trials. But these arguments overlook a more fundamental question: even if we could litigate all of the pending cases individually, should we? At first glance, the answer seems an obvious "yes." The apparent problem with sampling is that it makes extrapolations from the sampled cases to the nonlitigated cases, which introduces error. If one were to litigate all of the cases individually, however, this extrapolation error would disappear.

But extrapolation is not the only source of error when estimating damages. The jury itself is an imperfect device for measuring damages that produces error—at least in the statistical or scientific, rather than the legal, sense. On this score, sampling has distinct advantages. With individualized assessments, each case gets one jury, and absent remittitur or appellate reversal, the system is stuck with the result. Variabili-

---

[11] Two caveats to the discussion: First, for partly pedagogical and partly practical reasons, I will use damage estimation as the principal vehicle to discuss sampling. The dollar amounts in damage estimation are more illustrative than the dichotomous determinations in liability or causation, even though they are conceptually equivalent. Furthermore, from a practical standpoint, sampling liability would require the legal system to accept probabilistic notions of liability, something it has been loathe to do. Since liability and causation are arguably already more susceptible to class action or other aggregate treatment, sampling's principal venue will often be damages anyway. *See* Bone, *supra* note 10, at 597 (suggesting that under the current tort regime, sampling cannot be applied to liability because tort law does not "recognize[] probabilistic liability measures").

Second, the term "accuracy" in the context of damages concededly hides a fundamental controversy over whether certain kinds of noneconomic damages are capable of monetization. The typical criticism is that there is no such thing as "accuracy" because damages are socially constructed and there is no hard "truth" to be found. For the sake of brevity, I sidestep this philosophical debate. Since the legal system does indeed monetize noneconomic damages, I assume that there is some abstract value that litigation attempts to estimate. The fact that one can never directly measure or know the true value of damages—as is frankly the case with many real-world problems—does not prevent attempts to estimate it.

ties in the jury pool, mistakes made by the jury or the attorneys, or even nonjury related contingencies all become an unmitigated part of the litigation outcome. With sampling, however, case-specific contingencies even out because the case results are averaged across the sample. In short, sampling may introduce extrapolation error, but it also reduces variability.

From an accuracy standpoint, whether one prefers sampling or individualized trials is thus a function of case homogeneity and jury variability, an observation first made by Saks and Blanck.[12] If the sampled cases are very similar (which means low extrapolation error) or juries are very flaky (which means high measurement error), then sampling and averaging will produce more stable and accurate damage assessments than case-by-case adjudication. On the flip side, if the sampled cases are appreciably different, or juries are reliable, then the conventional preference for case-by-case adjudication holds. The desirability of sampling thus rests on two empirical questions: first, a general social science question about jury behavior and reliability; and second, a litigation-specific question about the homogeneity of the cases involved.[13]

In principle, the averaging advantage discussed above has very little to do with sampling per se. After all, one could reap the benefits of averaging merely by trying each case to multiple juries and averaging the results. But the working baseline is one that offers each party at

---

[12] *See* Saks & Blanck, *supra* note 10, at 833-37 (describing in detail the effect of sampling on reducing variation in case outcomes).

[13] The most serious critique of this analysis comes from Robert Bone, who argues that cases are rarely homogenous and that sampling may reduce attorney incentives, thereby negatively impacting accuracy. *See* Bone, *supra* note 10, at 576-94. The situation, however, may not be as grim as he suggests. While certain sampling or aggregation schemes can negatively impact attorney incentives, the problem can be minimized by pooling litigation costs and outcomes, a solution that Bone himself suggests. *Id.* at 591-92.

The heterogeneity concern requires some parsing. Some types of damages (often economic) are definitely heterogeneous—for example, lost wages, property damage, and medical expenses. Administrative measurement in these contexts, however, is arguably cheap, allowing the court to use regression modeling to fine-tune payouts after the sampling trials. Other damages, like pain and suffering, are more difficult to measure, but it is arguably appropriate to measure these amorphous damages using crude categories and assuming a homogeneous population. For example, pain and suffering resulting from paralysis or the loss of a limb should result in a standard predetermined sum, regardless of the other specifics of the case.

Interestingly, Bone's argument about attorney incentives potentially cuts in favor of sampling. If attorney incentives can significantly affect case outcomes, then there is likely to be a high variance in jury assessments, because their verdicts depend on attorney performance. But if juries are highly variable, then sampling will remain preferable even if the cases are not strictly homogeneous.

most *one* opportunity to litigate his case in full. Under those con-
straints, whether sampling or individual adjudication produces more
accurate results depends on the homogeneity of the group of cases
and an empirical question about jury behavior.

II. SHRINKAGE

Conventional thinking about mass litigation places a strong em-
phasis on commonality. Groups of litigants may be brought together
in a single proceeding—whether through a class action or multidistrict
litigation (MDL)—but generally only if common issues and interests
prevail among the group.[14] The theory behind commonality is quite
straightforward: group apples with apples, and separate apples from
oranges. Most aggregation procedures thus contemplate bringing liti-
gants together to resolve common issues, and then breaking them up
for separate litigation of party-specific issues.

Many of the cases involving sampling, however, do not adhere as
strongly to the commonality touchstone, again perhaps because of the
economic efficiencies required. For example, in *Cimino v. Raymark
Industries, Inc.*, Judge Parker sampled 160 asbestos cases from the 2298
on his docket, but the cases were far from homogeneous.[15] Indeed, he
ultimately divided the sample verdicts into five rather disparate cate-
gories, each corresponding to a particular asbestos-related disease.[16]

Beyond simply reducing costs, this kind of aggregate sampling also
turns out to potentially increase accuracy. Efron and Morris's classic
article about the Stein Paradox provides a useful starting point to
understand why this is so.[17] Suppose it is still early in the baseball sea-
son, and we want to estimate the likely batting average for a specific
player at the end of the season. Assuming no data from past years is
available, what is the most natural estimator of the player's end-of-
season average? Of course: the player's current batting average. In-

---

[14] *See* FED. R. CIV. P. 23(a)(2) (allowing class actions to proceed where "there are
questions of law or fact common to the class there are questions of law or fact common
to the class"); *see also* 28 U.S.C. § 1407 (2006) (providing for coordinated or consolidat-
ed pretrial proceedings when "one or more common questions of fact are pending in
different districts").

[15] *See* 751 F. Supp. 649, 653 (E.D. Tex. 1990) (describing a system for determining
damages in which sample cases from each of five disease categories went before a jury
and all other class members received the average award of those sample cases accord-
ing to their disease category), *aff'd in part, vacated in part*, 151 F.3d 297 (5th Cir. 1998).

[16] The five disease categories were mesothelioma, lung cancer, other cancer, asbes-
tosis, and pleural disease. *Id.*

[17] Bradley Efron & Carl Morris, *Stein's Paradox in Statistics*, SCI. AM., May 1977, at 119.

deed, one can show that current batting average is optimal in a statistical sense.[18] Now suppose that we want to estimate the season-end batting average for a larger number of players. Shouldn't we simply reapply the principle and use each player's current batting average to estimate his end-of-season average?

For years, statisticians thought precisely along these lines until Charles Stein surprisingly showed otherwise.[19] Stein showed that if one is making three or more estimates, considering information across players can result in a better estimate of an individual player's batting average.[20] The other players' statistics may initially appear irrelevant because Player *A*'s performance would seem to have no bearing on Player *B*'s performance. Additional reflection, however, suggests why Stein's discovery makes intuitive sense. Having data from the other baseball players gives us information about the league average and a player's performance relative to his peers. If a player is doing well early in the season, we might attribute this success to his ability, but we might also attribute it to random variation. After all, given enough baseball players, one expects that some will be lucky or unlucky early in the season, and that these cases will "regress to the mean" as the season wears on. Only a few exceptional players will distinguish themselves throughout the season. Consequently, adjusting or "shrinking" each individual's current batting average toward the overall average will result in estimates with lower overall error.[21]

Consider how case sampling procedures can thus improve the accuracy of litigation. In traditional litigation, the individual trial creates a single, isolated estimate of the plaintiff's damages. Sampling with strict commonality requirements can partially improve upon conven-

---

[18] Optimality here is defined in the classical sense of minimizing expected squared error.

[19] *See* Charles Stein, *Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution* (mathematically proving the inadmissability of the usual estimator), *in* 1 PROCEEDINGS OF THE THIRD BERKELEY SYMPOSIUM ON MATHEMATICAL STATISTICS AND PROBABILITY 197, 199-200 (1956); *see also* BRADLEY EFRON, LARGE-SCALE INFERENCE: EMPIRICAL BAYES METHODS FOR ESTIMATION, TESTING AND PREDICTION 1-12 (2010) (discussing James-Stein estimators generally).

[20] *See* Stein, *supra* note 19, at 199-200.

[21] The effectiveness of shrinkage turns out to apply even when one uses completely unrelated data. This "James-Stein Paradox" is the subject of various commentaries, although they are not applicable here. *See, e.g.,* Stephen M. Stigler, *The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators,* 5 STAT. SCI. 147, 148-49 (1990) (recasting the James-Stein estimator as a regression problem). The key insight for our purposes is that an excessive focus on individualized determinations may be suboptimal in the long run. Considering evidence about other, seemingly unrelated, individuals or individual cases can improve the overall accuracy of results.

tional practice, because sampling takes advantage of the aforementioned averaging effects. But further advantage comes from having a single decisionmaker consider several groups of related but not necessarily identical cases, much as Judge Parker did in the asbestos cases. This lumping of diverse groups, however, arguably violates the commonality required by current aggregation procedures.

Notably, the potential for aggregate sampling to achieve shrinkage-like results depends critically on the jury. If a single jury considers several disparate plaintiff populations simultaneously, will the jury make assessments that are consistent with a statistical shrinkage estimator? The question is again an empirical one, but intuitively, one imagines that juries might perform at least some shrinkage when provided with multiple populations. For example, in *Cimino*, having five different asbestos-related diseases under consideration presumably helped the jury calibrate its assessment of damages in each case.[22] Similarly, baseball fans automatically do shrinkage when relying on their prior experience with the game. When a player bats a phenomenal .500 during the month of April, no one thinks that .500 is going to be the player's season-end average; surely it will come down over time.

The most concerning aspect of shrinkage is that while it reduces aggregate error, that reduction comes at the potential cost of increased error in individual cases.[23] In the baseball example, the shrinkage estimates for truly exceptional players may exhibit large errors, since the procedure cannot distinguish talent from chance variation. So although shrinkage reduces the overall (league-wide) error, predictions for star players may be more inaccurate than if we had considered each player individually. Analogously, while shrinkage may reduce overall systemwide error in mass tort damage assessments, plaintiffs with exceptionally large damages may suffer more error than under individualized adjudication. We of course do not know which specific litigants (or baseball players) are worse off. We only know that the risk exists.

This tradeoff between systemwide error and error in the individual case poses an important question: what kind of error should the legal system be minimizing? Is it the expected overall error across cases or the expected error in each individual case? As the shrinkage case illustrates, the two are not always commensurate—optimizing for one

---

[22]   *See supra* note 16 and accompanying text.

[23]   In statistical terms, shrinkage estimators do not *uniformly* reduce expected error.

property may come at the cost of the other.[24] The knee-jerk response may be that minimizing error in the individual case is paramount, since it would be unfair for one litigant to "pay" for the better results of the whole. But we make this kind of tradeoff in evidence law all the time. Evidentiary rules are set up in the hope of minimizing overall error across cases, even if the rules may harm some (unidentifiable) cases as a result.

III. NONRANDOM SAMPLING

One final—and striking—aspect of court-imposed sampling is that such sampling is often nonrandom. For example, the six cases ultimately tried in the *In re World Trade Center Disaster Site Litigation* were hand-picked: two by the plaintiffs, two by the defense, and two by the judge.[25] At first glance, this kind of adversarial sampling seems unscientific, a bit lazy, and perhaps even dangerous, with the only acceptable justification being the need to secure party consent. The dangers of convenience sampling are well known, and since the entire case population was available to the court, failure to use random sampling seems inexcusable.[26]

Here again, however, sampling procedures—in this case, *nonrandom* sampling procedures—can under certain circumstances counterintuitively increase accuracy. Generally speaking, nonrandom sampling is undesirable because it introduces bias: it tilts the estimates in specific (although not always known) directions. But the nonrandom sampling seen in mass tort cases is a very particular kind. By relying on party selection, courts effectively sample from the extremes of the distribution, which under some conditions can result in better estimates than randomly sampling from the whole. Broadly speaking, party-selected sampling can take advantage of information held by the parties to construct a more efficient sampling procedure.

---

[24] Statistical decision theory considers further, more nuanced properties for statistical estimators. For example, when choosing an estimator, one may also want it to minimize the maximum possible error (minimaxity) or find the exactly correct answer given unlimited data (consistency), and so forth. These complexities are beyond the scope of the present discussion.

[25] Alexandra D. Lahav, Rough Justice 17 (Aug. 9, 2010) (unpublished manuscript), *available at* http://ssrn.com/abstract=1562677; *see also id.* at 18 (dubbing this process "court engineered sampling").

[26] *See id.* at 24-25 (criticizing courts for not using random sampling since such techniques are not likely to provide reliable information); *see also* Saks & Blanck, *supra* note 10, at 841-42 (characterizing mass torts as a "sampling theorist's dream" since the completeness of the population enables excellent random sampling).

Extreme value sampling apparently has ancient roots. In a recent article, Davis, Friedman, and Ye discuss its appearance in the rabbinical literature associated with the rather mundane (but apparently then important) problem of estimating the volume of chicken eggs.[27] As they note, while averaging extreme values may initially seem "seriously flawed," under certain conditions, it is "unexpectedly good, and . . . may even [be] optimal."[28] The problem with extreme value averaging is that it is highly nonrobust, meaning it is badly sensitive to outliers and performs terribly on asymmetric distributions.[29] However, if the quantity being measured comes from a symmetric distribution with minimal outliers (e.g., a normal distribution), then this method can be superior to random sampling. For example, Davis, Friedman, and Ye show that when the target is normally distributed and the population is large (five hundred cases), randomly selecting one case from the top ten percent of the population, another from the bottom ten percent, and averaging the results is statistically equivalent to a sample size of twelve from the full population.[30] So rather than conduct twelve trials, the court need only conduct two. As one might expect, further benefits accrue if the court tries additional cases from the extremes of the population, although with diminishing returns.

Concededly, there are good reasons to avoid extreme value sampling, particularly since we often do not have a good sense of the underlying distribution and cannot guarantee that the necessary conditions exist.[31] There are, however, more sophisticated and better accepted nonrandom sampling schemes, and these may be suitable replacements for the current one.[32] But for our purposes, the statistical spe-

---

[27] Harry Zvi Davis, Hershey H. Friedman & Jianming Ye, *An Ancient Sampling Technique: Flawed, Surprisingly Good, or Optimal?*, CHANCE, Mar. 2011, at 19, 19-20.

[28] *Id.* at 19.

[29] Indeed, for a chi-square distribution, the larger the sample, the *worse* the estimate generated by extreme value averaging. *See id.* at 20.

[30] *Id.* at 22 tbl.3 (showing that the mean of a random selection from among the largest eggs and a random selection from among the smallest eggs is the same as the result from 11.5 sampled eggs from a batch of 500, or 11.7 from a batch of 1000).

[31] Some may worry that extreme value averaging will encourage plaintiffs to bring frivolous cases, but the incentives actually suggest the opposite result. Since defendants get to choose a subset of the tried cases, bringing frivolous cases only helps defendants. Plaintiffs thus have significant incentives to bring only good claims.

[32] For example, stratified sampling techniques can capture some of the gains of extreme value sampling while mitigating some of the dangers. Further, if one really wanted to maximize effective sample size for estimating the mean, the median would be a terrific estimator (although it is unclear how one could estimate the median without full litigation). *See generally* Gang Zheng & Joseph L. Gastwirth, *Where Is the Fisher*

cifics are beside the point. The point is that, once again, considering the population as a whole can help achieve better estimates than considering each case as an island. In this case, having the two sides effectively order their cases from best to worst provides insights that would otherwise be lost.

CONCLUSION

This Essay has offered a line of accuracy-based arguments in favor of sampling.[33] Despite its origins as a mechanism to improve economic efficiency, trial sampling is not a "second-best" option to be considered only because individual adjudication is economically impractical or impossible. To the contrary, sampling has unexpected advantages in averaging, shrinkage, and information gathering that can make it preferable to individualized adjudication, regardless of what our intuitions might initially suggest.

Although the focus of this discussion has been on mass torts, the analysis has implications that extend further: if sampling is not "second-best," then the legal system arguably should consider it *even when* it has the resources to litigate individual cases. For example, would damage assessments in recurring cases such as medical malpractice or car accidents benefit from some form of aggregation and sampling? Because the underlying facts in these cases vary considerably, the exact mechanism for sampling may be complicated, but as this Essay suggests, nothing inherently says that individual trials are always the way to go.

Information in an Ordered Sample?, 10 STATISTICA SINICA 1267, 1275 (2000) (discussing where most of the information about the mean comes from in an ordered sample).

[33] *Cf.* Lahav, *supra* note 25, at 29-30 (arguing that "sampling is desirable not only because it is *efficient* but also because it is *fair*").

* * * * * *