

When 10 Trials Are Better Than 1000: An Evidentiary Perspective on Trial Sampling

Edward K. Cheng*

* Professor of Law, Vanderbilt Law School, and doctoral student, Department of Statistics, Columbia University. Thanks to Joe Gastwirth, Jenny Diamond Cheng, Robin Efron, Chris Robertson, Magda Hanebach, and the students of my Statistical Inference and the Law seminar at Brooklyn Law School for helpful thoughts and suggestions. This paper is dedicated to my late colleague Richard Nagareda.

In many mass tort cases, individual trials are simply impractical. Take for example *Dukes v. Wal-Mart Stores, Inc.*,¹ a class-action employment discrimination suit currently before the Supreme Court. With over 1.5 million women in the *Dukes* plaintiff class,² the notion of individual trials is so ridiculous as to induce smiles. Other recent notable examples of the phenomenon include the World Trade Center Disaster Site litigation³ and the fraud litigation against light cigarette manufacturers, in which Judge Weinstein colorfully noted that any “individualized process . . . [would] continue beyond all lives in being.”⁴

Faced with an unserviceable number of plaintiffs, courts have proposed sampling trials – rather than litigating each and every case, the court would litigate a small subset of the claims and award the remaining litigants statistically determined amounts based on the results. But while sampling is standard statistical practice and often accepted as evidence in other legal contexts,⁵ appellate courts have balked at the notion of court-mandated, binding trial sampling over due process concerns.⁶

¹ 603 F.3d 571 (9th Cir. 2010), *cert. granted in part*, 131 S.Ct. 795 (2010).

² *Dukes v. Wal-Mart Stores*, 603 F.2d 571, 628-29 (9th Cir. 2010) (Ikuta, J., dissenting).

³ Order Amending Case Management Order No. 8, *In re World Trade Center Disaster Litigation*, 21 MC 100 (Feb. 19, 2009).

⁴ *Schwab v. Philip Morris USA, Inc.*, 449 F. Supp. 2d 992, 1018 (E.D.N.Y. 2006), *rev'd*, *McLaughlin v. American Tobacco Co.*, 522 F.3d 215 (2d Cir. 2008).

⁵ *E.g.*, Shari Seidman Diamond, *Survey Evidence*, in 1 David L. Faigman, et al., *Modern Scientific Evidence*, § 8.2, at 483-84 (2009-10 ed.) (discussing the use of sampling surveys in trademark litigation); *see also* 2 Joseph L. Gastwirth, *Statistical Reasoning in Law and Public Policy* [jumpcite] (describing other uses of sampling in the legal system).

⁶ *E.g.*, *McLaughlin*, 522 F.3d at 231 (rejecting trial court’s proposal to use sample trials to determine aggregate liability in light cigarette litigation); *Cimino v. Raymark Indus.*, 151 F.3d 297, 319-20 (5th Cir. 1998) (rejecting use of sampling in asbestos cases). *But see* *Hilao v. Estate of Marcos*, 103 F.3d 767, 782-87 (9th Cir. 1996) (noting that the sampling “methodology in determining valid claims [was] unorthodox, . . . [but could] be justified by the extraordinarily unusual circumstances”); *but see also* *Dukes*, 603 F.3d at 624-25 & n.53 (noting exception to preference for individual hearings when individualized evidence is difficult to obtain). *See generally* Laurens Walker & John Monahan, *Sampling Evidence at the Crossroads*, 80 So. Cal. L. Rev. 969 (2007) (discussing Judge Weinstein’s proposed use of sampling in the *McLaughlin* case, among other things).

Despite appellate reluctance, the controversy nonetheless rages on. Trial courts have solidified on through non-binding sampled trials (dubbed “bellwether trials”) to induce settlement,⁷ and recent courts like the Ninth Circuit in *Dukes* have hinted at greater receptivity to the concept.⁸ Given that trial courts have few practical alternatives, one wonders if it is just a matter of time before their appellate brethren recognize sampling’s necessity.

As one might expect, the most common and salient argument for trial sampling is economic efficiency. Since the legal system lacks the resources to litigate hundreds of thousands of asbestos cases, some kind of resolution of these cases seems better than none. Otherwise, the tort system would face a profound loss of deterrence against wrongdoers and compensation to victims. Opponents’ objections predictably take a liberty or rights-based approach: Defendants are simply entitled to individual trials; approximate justice will not do, no matter what the social costs.

These debates over efficiency and individual rights are wholly consistent with sampling’s origins in civil procedure. In contrast, this Essay develops the evidentiary perspective on the sampling controversy, a perspective that has received relatively less attention.⁹ Putting aside economic and liberty interests, what effect does sampling actually have on *accuracy*? Implicit in most discussions is the assumption that sampling is a “second best” solution, contemplated only in the most dire of circumstances. Individual trials are the preferred basis of the legal system – a basis not without flaws, but presumably the best we can do. So if we could try all 1.5 million cases in *Dukes*, wouldn’t we want to? Surely since sampling involves *estimating* liability from a selected subset of cases, it is suboptimal compared to traditional, individualized adjudication. Or is it?

In pages that follow, I offer three ways in which this “second best” assumption can be wrong. Sampling is not necessarily second best, and given the right conditions, it can actually be superior to individualized adjudication in determining case outcomes. Intuitively, sampling’s advantage comes from its ability to borrow

⁷ Alexandra Lahav, *Bellwether Trials*, 76 Geo. Wash. L. Rev. 576, 581 (2008).

⁸ *Dukes*, 603 F.3d at 625-26 (expressing “no opinion” about the district courts specific trial plan, but showing affinities for sampling plans).

⁹ Two earlier cross-cutting discussions are found in Michael J. Saks & Peter David Blanck, *Justice Improved: The Unrecognized Benefits of Aggregation and Sampling the Trial of Mass Torts*, 44 Stan. L. Rev. 815 (1992), and Robert G. Bone, *Statistical Adjudication: Rights, Justice, and Utility in a World of Process Scarcity*, 46 Vand. L. Rev. 561 (1993).

strength across the different cases in the sample. Individual adjudication narrowly confines itself to a single case and factfinder; sampling does not. This basic principle recurs in the following sections on averaging, shrinkage, and non-random sampling. What's more, the good news is that the sampling procedures proposed by courts frequently capture these advantages, even if their original impetus may have been reducing cost, not improving accuracy.¹⁰

1 Averaging

As noted in the Introduction, the primary motivation for sampling cases is to reduce litigation costs. Sampling in asbestos, the World Trade Center litigation, light cigarettes, or any other major mass tort is necessary because courts are simply not equipped to run hundreds of thousands of trials. But these arguments overlook a more fundamental question - in terms of accuracy, if we could litigate all of the pending cases individually, should we? At first glance, the answer seems an obvious "yes." The big problem with sampling is that it extrapolates from the sampled cases to the non-litigated cases, and this extrapolation creates error. If one litigates all of the cases individually, the extrapolation error disappears.

Extrapolation, however, is not the only source of error when trying to estimate damages. If one thinks of a jury as an (imperfect) device for measuring damages, then it too produces error (in the statistical or scientific, rather than legal sense). On this score, sampling has distinct advantages. With individualized assessments,

¹⁰ 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 statistically 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, and since liability and causation are arguably more susceptible to class-action or other aggregate treatment, sampling's principal venue will often be damages anyway. See Bone, *supra* note 9, at 597.

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 tries to estimate. The fact that one can never directly measure or know the true value does not prevent attempts to estimate it, as is frankly the case with most statistical problems.

each case gets one jury, and absent remittitur or reversal, the system is stuck with the result. Variabilities in the jury pool, mistakes made by the jury or the attorneys, or even non-jury-related contingencies – all become an unmitigated part of the litigation outcome. With sampling however, case-specific contingencies even out, because as each case in the sample is litigated, the results are averaged with other cases in the sample. In short, sampling may introduce extrapolation error, but it reduces variability.

From an accuracy standpoint, whether one prefers sampling or individualization is thus a function of case homogeneity and jury variability, an observation that was first made by Saks and Blanck.¹¹ If the sampled cases are very similar (which means low extrapolation error) or juries extremely flaky, then sampling and averaging will produce more stable and accurate damage assessments than relying on 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 cases.¹²

Notably, 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 most *one* opportunity to litigate his/her case in

¹¹ Saks & Blanck, *supra* note 9, at 833-37.

¹² The most serious critique of this analysis comes from Robert Bone, who argues that cases are rarely homogenous and that sampling may disincentivize attorneys and thereby negatively impact accuracy. See Bone, *supra* note 9, at 576-94. The situation, however, may not be as dire as he suggests. While certain sampling or aggregation schemes can negatively impact attorney incentives, pooling litigation costs and outcomes can minimize the problem, as Bone himself suggests. *Id.* at 590-92. The heterogeneity concern requires some parsing. Some types of damage (usually economic) are definitively heterogenous – for example, lost wages, property damage, and medical expenses. But administrative measurement in these contexts is relatively cheap, allowing the court to use regression modeling to fine-tune payouts after the sampling trials. Noneconomic damages such as pain-and-suffering are difficult to measure, but being amorphous, they are arguably appropriately modeled as arising from homogenous populations (e.g., loss of hand, paralysis, etc.).

Interestingly, the concern over attorney incentive effects somewhat cuts against the concern about case heterogeneity. If attorney incentives can significantly affect case outcomes, then we basically have high variance in jury assessments. But if juries are highly variable, then sampling will remain preferable even if the cases are not strictly homogeneous.

full. Under those constraints, whether sampling or individual adjudication is optimal depends on the characteristics of the group of cases and an empirical question about jury behavior.

2 Shrinkage

Conventional thinking about mass litigation places a strong emphasis on commonality. Groups of litigants may be brought together in a single litigation – whether a class action, multidistrict litigation (MDL), or even joinder – but only if common issues and interests prevail among the group. The theory behind commonality is quite straightforward: treat apples with apples, and separate apples from oranges. Most aggregation procedures thus contemplate bringing litigants 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*, Judge Parker sampled 160 asbestos cases from the 2300 on his docket, but the cases were far from homogeneous.¹³ Indeed, he ultimately divided the sample verdicts into five rather disparate categories, each corresponding to a particular asbestos-related disease.¹⁴

Far from simply reducing costs, however, this kind of aggregate sampling turns out to potentially increase accuracy. To understand how having a single decisionmaker share or mix information across different cases or classes of litigants might improve accuracy, Efron and Morris's classic article about the Stein Paradox provides a useful starting point:¹⁵ Suppose it is still early in the baseball season. Assuming data from past years are not available, we want to estimate the likely batting average for a specific player at the end of the season. What is the most natural estimator? Of course, the player's current batting average – and indeed, one can show that current batting average is optimal in a statistical sense.¹⁶ Now suppose that we want to estimate the season-end batting averages for

¹³ *Cimino v. Raymark Industries*, 751 F. Supp. 649, 653 (N.D. Tex. 1990).

¹⁴ *Id.* at 664-65.

¹⁵ Bradley Efron & Carl Morris, *Stein's Paradox in Statistics*, 236 *Scientific American* 119 (1977).

¹⁶ Optimality here is defined in the classical sense of minimizing expected squared error.

a whole group of players. Shouldn't we just extend the idea and use the current batting averages for each player?

For years, statisticians thought precisely that, until Charles Stein surprisingly showed otherwise.¹⁷ 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. The other player's stats may initially seem totally irrelevant, because Player A's performance would seem to have no bearing on Player B's, but additional reflection 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 notably well early in the season, we might attribute the success to player 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 exceptional players will keep it up throughout the season. Consequently, adjusting or "shrinking" each individual's current batting average toward the overall average results in estimates with lower overall error.¹⁸

Consider how case sampling procedures can thus improve the accuracy of litigation. In conventional litigation, the individual trial creates a single, isolated estimate of the plaintiff's damages. Sampling with strict commonality requirements can be an improvement, because it 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. As noted above, this lumping of diverse groups arguably violates the commonality required by most aggregative procedures. Practically speaking, the aggregation is appealing because it is efficient—it avoids relitigation of the baseline underlying issues, such as wrongful conduct or causation. But because of the shrinkage

¹⁷ C. Stein, *Inadmissibility of the Usual Estimator for the Mean of a Multivariate Distribution*, 1 Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 197 (1956); see also Bradley Efron, *Large-Scale Inference* 1-12 (2010) (discussing James-Stein estimators generally).

¹⁸ 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, but is largely superfluous for our purposes. The key insight is that excessively focusing on individualized determinations may be suboptimal in the long run. Considering evidence about other, seemingly unrelated, people can improve the overall accuracy of results.

phenomenon, broader aggregation can also lead to lower overall error.

Notably, the ability for aggregative sampling to achieve shrinkage-like results depends critically on the jury. When considering several disparate plaintiff populations, does a jury make assessments in a way consistent with a statistical shrinkage estimator? The question is of course an empirical one, but intuitively, one imagines that juries might do 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 for each disease or case. Recalling the baseball average example, 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 the best prediction for 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.¹⁹ 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 while shrinkage reduces the overall error, its predictions for star players may be worse than if we had only considered each player individually. Similarly, in the mass tort context, while shrinkage may reduce overall error in damage assessments, people with outlying damages may suffer more error than under individualized adjudication. We of course do not know which specific estimates are worse off, since we are ignorant of the truth. We just know that the risk exists.

This tradeoff between systemwide error and error in the individual case raises the important question of just what kind of error the legal system should be minimizing. Is it the expected overall error across cases? Or is it the expected error in each individual case? As the shrinkage case shows, the two are not always commensurate—optimizing for one property may come at the cost of the other.²⁰ The knee-jerk response may be that error in

¹⁹ In statistical terms, shrinkage estimators do not *uniformly* reduce expected error.

²⁰ Statistical decision theory asks even more nuanced properties of statistical estimators – for example, that they minimize the maximum error that can be experienced (minimax), or that given unlimited data, they find the exactly correct

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 tradeoff in evidence law all the time. Evidentiary rules are set up in the hope of minimizing overall error across cases, even if they may prospectively harm some cases as a result.

3 Non-Random Sampling

One final, striking aspect of court-imposed sampling is that it is often non-random. For example, the six cases ultimately tried in the World Trade Center Disaster Site litigation were hand-picked: two by the plaintiffs, two by the defense, and two by the judge.²¹ At first glance, this kind of adversarial sampling seems unscientific, perhaps even lazy, 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 is available to the court, failure to use random sampling seems inexcusable.²²

Here again however, the sampling procedure used by courts may, under certain circumstances, counterintuitively increase accuracy. Generally speaking, non-random sampling is undesirable because it introduces bias – it tilts the estimates in specific (although not always known) directions. But the non-random sampling seen in mass tort cases is a very particular kind of non-random sampling. 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.

Extreme value sampling apparently has ancient roots. In a recent article, Davis, Friedman & 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.²³ As they note, averaging extreme values may initially seem “seriously flawed,” but under certain conditions, it is “unexpectedly good, and .

answer (consistency), and so forth. These complexities are beyond the scope of the present discussion.

²¹ Alexandra Lahav, *Rough Justice*, Aug. 9, 2010 (unpublished manuscript), at 17.

²² Lahav, *supra* note 21, at 24 (criticizing courts for not using random sampling); Saks & Blanck, *supra* note 9, at 841-42 (characterizing mass torts as a “sampling theorists’ dream” since the completeness of the population enables excellent random sampling).

²³ Harry Zvi Davis, Hershey H. Friedman & Jianming Ye, *An Ancient Sampling Technique: Flawed, Surprisingly Good, or Optimal?*, 24 *Chance* 19 (2011).

. . . may even [be] optimal.”²⁴ The problem with extreme value averaging is that it is highly “non-robust,” meaning that it is badly sensitive to outliers, and it performs terribly on non-symmetric distributions.²⁵ However, if the quantity being measured comes from a symmetric distribution with minimal outliers (e.g., a normal distribution), then the method can be superior to random sampling. For example, Davis et al. show that when the target is normally distributed and the population is large (500 cases), randomly selecting a case from the top 10% and bottom 10% of the population and averaging the results is effectively the same as having a sample size of 12.²⁶ 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.

Now, as just acknowledged, 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. There are, however, more sophisticated and better accepted non-random sampling schemes, and these may be suitable replacements for the current one.²⁷ But for our purposes, the sampling specifics 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 order their cases provides insights that would otherwise be lost.

²⁴ *Id.* at 19.

²⁵ Indeed, for a chi-square distribution, the larger the sample, the *worse* the estimate generated by extreme value averaging. *Id.* at 20.

²⁶ *Id.* at 22 tbl.3.

²⁷ For example, there are stratified sampling techniques, and 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 Zhang & Joseph L. Gastwirth, *Where is the Fisher Information in an Ordered Sample?*, 10 *Statistica Sinica* 1267, 1275 (2000) (discussing where most of the information about the mean comes from an ordered sample).

4 Conclusion

This Essay has offered a new line of accuracy-based arguments in favor of sampling.²⁸ Despite having its origins in 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 say.

Although the focus of this discussion has been on mass torts, the analysis has implications that extend beyond, for if sampling is not “second-best,” then arguably the legal system 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 would vary considerably, the exact mechanism might be complicated, but as this Essay suggests, nothing inherently says that individual trials are always the way to go.

²⁸ Cf., e.g., Lahav, *supra* note 21, at 29-49 (developing arguments that sampling has fairness implications).